Robert Geroch

Lectures on

Mathematical Methods in Physics

Compiled by

Sati Shankar

orcid.org/0000-0003-4638-1745

from

https://uchicago.app.box.com

2020

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

To the Readers

During my varied multidisciplinary studies, especially in my transition from mathematical economics to theoretical physics, I found the style of exposition of Prof. Geroch more intuitively and seamlessly formalised than many other such sources.

Sitting at my place I reached these lectures the harder way. It is this fact which had led me compile scattered material with a view to help others, especially future physicists and theoretical researchers.

I have just compiled the scattered study materials, all the IPR related issues and other rights, attached to respective documents compiled herewith, shall be as they have been with the original authors and publishers. Any unlawful use of this compilation must be avoided.

Best wishes

Sati Shankar

orcid.org/0000-0003-4638-1745

GSFN, New Delhi, India

satigsf at gmail dot com

Table of Content

1	**Geometrical Quantum Mechanics**	**1-73**
2	**Differential Geometry**	**74-175**
3	**Infinite Dimensional Manifolds**	**176-263**
4	**General Relativity**	**264-393**
5	**Integration Theory**	**395-489**
6	**Topology**	**490-607**
7	**Special Topics in Physics**	**608-891**
8	**Perspectives in Computation**	**729-891**
9	**Unsolvable problems**	**892-966**

GEOMETRICAL QUANTUM MECHANICS

Robert Geroch (University of Chicago, 1974)

TEXed for posterity by a grad student from an n^{th}-generation photocopy of the original set of lecture notes. (Aug 1994)

transcribed version: 1/31/2006

Contents

I. Differential Geometry 1

1. Manifolds 1

2. Tensor Algebras 4

3. Tensor fields 7

4. Derivative Operators 10

II. Mechanics 13

5. Configuration Space 13

6. The Cotangent Bundle 18

7. Symplectic Manifolds 20

8. Phase Space 22

9. The Hamiltonian 23

10. Observables; Poisson Bracket 25

11. Canonical Transformations 27

12. Algebraic Observables 31

III. Quantum Mechanics 33

13. Introduction 33

14. Densities. Integrals. 35

15. States 38

16. Observables 40

17. Higher Order Observables 44

18. The Hamiltonian. Dynamics 47

19. Higher Order Observables—Revisited 49

20. Heisenberg Formulation 51

21. The Role of Observables. (Classical Mechanics) 56

22. The Role of Observables. (Quantum Mechanics) 59

23. The Interpretations of Quantum Mechanics 63

24. The Copenhagen Interpretation 65

25. The Everett Interpretation 67

GEOMETRICAL QUANTUM MECHANICS

Robert Geroch (University of Chicago, 1974)

TEXed for posterity by a grad student from an n^{th}-generation photocopy of the original set of lecture notes. (Aug 1994)

Part I.
Differential Geometry

1. Manifolds

The arena in which all the action takes place in differential geometry is an object called a *manifold*. In this section, we define a manifold, and give a few examples.

Roughly speaking, a n-dimensional manifold is a space having the "local smoothness structure" of Euclidean n-space, $\mathbb{R}^n$. (Euclidean n-space is the set consisting of n-tuples, $(x_1, \ldots, x_n)$, of real numbers). The idea, then, is to isolate, from the very rich structure of $\mathbb{R}^n$ (*e.g.*, its metric structure, vector-space structure, topological structure, etc.), that one bit of structure we call "smoothness".

Let M be a set. An n-**chart** on M consists of a subset U of M and a mapping $\psi : U \to \mathbb{R}^n$ having the following two properties:

1. The mapping ψ is one-to-one.

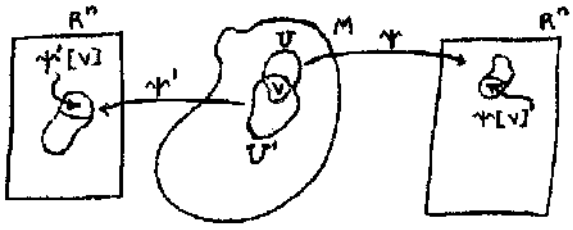
   (That is, distinct points of U are taken, by ψ, to distinct points of $\mathbb{R}^n$.)

2. The image of U by ψ, *i.e.*, the subset $O = \psi[U]$ of $\mathbb{R}^n$, is open in $\mathbb{R}^n$.

   (Recall that a subset O of $\mathbb{R}^n$ is said to be **open** if, for any point x of O, there is a number $\epsilon > 0$ such that the ball with center x and radius ϵ lies entirely within O.)

Let (U, ψ) be a chart. Then, for each point p of U, $\psi(p)$ is an n-tuple of real numbers. These numbers are called the **coordinates** of p (with respect to the given chart). The range of coordinates (*i.e.*, the possible values of $\psi(p)$ for p in U) is the open subset O of $\mathbb{R}^n$. By (1), distinct points of U have distinct coordinates. Thus, ψ represents n real-valued functions on U; a chart defines a labeling of certain points of M by real numbers.

These charts are the mechanism by which we intend to induce a "local smoothness structure" on the set M. They are well-suited to the job. Since a chart defines a correspondence between a certain set of points of M and a certain set of points of $\mathbb{R}^n$, structure on $\mathbb{R}^n$ can be carried back to U. To obtain a manifold, we must place on M a sufficient number of charts, and require that, when two charts overlap, the corresponding smoothness structures agree.



Let (U, ψ) and (U', ψ') be two n-charts on the set M. If U and U' intersect in M, there is induced on their intersection, $V = U \cap U'$, two "smoothness

structures". We wish to compare them. To this end, we introduce the mapping $\psi' \circ \psi^{-1}$ from $\psi[V]$ to $\psi'[V]$ and its inverse $\psi \circ \psi'^{-1}$ from $\psi'[V]$ to $\psi[V]$. But $\psi[V]$ and $\psi'[V]$ are subsets of $\mathbb{R}^n$, hence $\psi' \circ \psi^{-1}$ represents n functions-of-n-variables: $x'^1(x^1, \ldots, x^n),\ \ldots, x'^n(x^1, \ldots, x^n)$. Similarly, $\psi \circ \psi'^{-1}$ represents the inverse functions: $x^1(x'^1, \ldots, x'^n),\ \ldots,\ x^n(x'^1, \ldots, x'^n)$. These functions represent the interaction of (U, ψ) and (U', ψ') as regards smoothness structure on $V = U \cap U'$. We are thus led to the following definition: the n-charts (U, ψ) and (U', ψ') on M are said to be **compatible** if

1. $\psi[V]$ and $\psi'[V]$ are open subsets of $\mathbb{R}^n$
2. the mappings $\psi' \circ \psi^{-1}$ and $\psi \circ \psi'^{-1}$ are C^∞ (*i.e.*, all partial derivatives of these functions, of all orders, exist and are continuous).

Note that, for compatibility of charts, we require only that they agree in one structure of interest: *smoothness*. It is in this way that a single structure is isolated.

An **n-dimensional manifold** consists of a nonempty set M, along with a collection of n-charts on M such that

1. Any two charts in the collection are compatible.
2. Any chart on M which is compatible with all charts in the collection is also in the collection.
3. The charts in the collection cover M, *i.e.*, every point of M is in at least one of the charts.
4. The charts separate points of M, *i.e.*, if p and q are distinct points of M, then there are charts (U, ϕ) and (U', ϕ') in our collection such that p is in U, q is in U'; U and U' do not intersect.

These conditions—or at least the first three—are exactly what one might expect intuitively. The first condition states that *"whenever two charts induce competing smoothness structures on the same region of M, the structures agree"*. The second condition ensures that *we have "enough" charts*. The third condition ensures that *"smoothness structure is induced over all of M"*. The fourth condition eliminates pathological objects (called *non-Hausdorff* manifolds) which are of little interest.

One quickly gets an intuitive feeling for manifolds, so it becomes unnecessary constantly to return to these awkward definitions.

Example 1. Let M be the set $\mathbb{R}^n$, *i.e.*, the set of all n-tuples of real numbers. This M comes equipped already with a chart: set $U = M$, and ϕ the identity mapping (from M to $\mathbb{R}^n$). Now consider all n-charts on M which are compatible with this one. This set M, together with these charts, satisfies our four conditions for a manifold. (Condition 1 follows from the fact that smooth functions of smooth functions are smooth. The second and third conditions are obvious, and the fourth nearly so.) This n-dimensional manifold is called the manifold $\mathbb{R}^n$.

Example 2. Let M be the set of $(n+1)$-tuples of real numbers, $(y^1, \ldots, y^{n+1})$, satisfying $(y^1)^2 + \cdots + (y^n)^2 = 1$. Denote by U the subset of M consisting of $(y^1, \ldots, y^{n+1})$ with $y^1 > 0$. Let ϕ denote the mapping from U to $\mathbb{R}^n$ defined by $x^1 = y^2,\ \ldots,\ x^n = y^{n+1}$. This is a chart on M. Similarly, introduce charts given by $y^1 < 0,\ y^2 > 0,\ y^2 < 0,\ \ldots,\ y^{n+1} < 0$. In this way, we obtain $2n+2$ charts on M which, as can be easily checked, are compatible with each other. Now consider all charts on M which are compatible with these $2n+2$ charts. As in Example 1, one can verify that this set M, with these charts, defines a manifold. It is called the manifold S^n, (the n-sphere).

Example 3. Let M be the set of all orthogonal 3×3 matrices (*i.e.*, matrices whose inverse is equal to their transpose). Let U consist of the subset M consisting of matrices are within $\frac{1}{10}$ of the corresponding entries in the unit-matrix. That is, U consists of orthogonal matrices $\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$ with $|a-1|$, $|e-1|$, $|i-1|$, $|b|$, $|c|$, $|d|$, $|f|$, $|g|$, $|h|$ all less than $\frac{1}{10}$. For such a matrix K, set $\phi(K) = (b, c, f)$. This is a 3-chart on M. Fix any orthogonal matrix P. We define another chart, with U' consisting of matrices Q with QP in U, and with $\psi'(Q) = \psi(QP)$. Thus, for each P we obtain another chart. These charts are all compatible with each other. The collection of all charts on M compatible with these makes our set M into a 3-dimensional manifold. This is the manifold of the Lie-group $O(3)$. (Similarly, other matrix groups are manifolds).

2. Tensor Algebras

The natural things to exist on a manifold—the things which will describe the physics on a manifold—are objects called *tensor fields*. It is convenient to begin by ignoring manifolds completely. We shall define an abstract *tensor algebra*—a structure which makes no reference to a manifold. We shall then see, in the following section, that these tensor algebras arise naturally on manifolds.

A tensor algebra consists, first of all, of a certain collection of sets. We shall label these sets by a script "T" to which there is attached subscripts and superscripts from the Latin alphabet. In fact, we introduce a set for every arrangement of such subscripts and superscripts on $\mathcal{T}$, provided only that the arrangement is such that no index letter appears more than once. Thus, for example, our collection includes sets denoted by $\mathcal{T}^m{}_{rs}{}^b{}_a$, $\mathcal{T}_c$, $\mathcal{T}$, etc. The collection does not, however, include a set denoted by $\mathcal{T}_{dc}{}^{ac}{}_r$.

The elements of these sets will be called **tensors**. To indicate the particular set to which a tensor belongs, we shall write each tensor as a base letter with appropriate subscripts and superscripts. Thus, for example, $\xi_r{}^{pa}{}_{sd}$ represents a tensor, namely, an element of $\mathcal{T}_r{}^{pa}{}_{sd}$. Tensors with a single index (*e.g.*, η_c, an element of $\mathcal{T}_c$) are called **vectors**; tensors with no indices (e.g., α, an element of $\mathcal{T}$) **scalars**.

The next step in the definition of a tensor algebra is the introduction of four operations on these tensors.

1. Addition. With any two tensors which are elements of the same set (*i.e.*, which have precisely the same index structure), there is associated a third tensor, again an element of that same set. This operation will be written with a "+". Thus, $\lambda^c{}_{ad} + \omega^c{}_{ad}$ is an element of $\mathcal{T}^c{}_{ad}$.

2. Outer Product. With any two tensors having no index letter in common, there is associated a third tensor, an element of the set denoted by "$\mathcal{T}$" with first the indices of the first tensor, and then the indices of the second tensor attached. This operation will be indicated by writing the two tensors next to each other. Thus, for example, $\mu_{we}{}^r{}_t\nu^{yu}{}_c$ is an element of $\mathcal{T}_{we}{}^r{}_t{}^{yu}{}_c$.

3. Contraction. Given any tensor, along with a choice of a particular subscript and a particular superscript of that tensor, there is associated another tensor, an element of the set denoted by "$\mathcal{T}$" with all the indices of the original tensor, except the two chosen ones, attached. This operation is indicated by writing the original tensor with the chosen subscript changed so as to be the same letter as the chosen subscript. Thus, for example, choosing the superscript "c" and the subscript "m" for $\xi_r{}^d{}_{mas}{}^{bc}$, we obtain by contraction the element of $\mathcal{T}_r{}^d{}_{as}{}^b$ written $\xi_r{}^d{}_{mas}{}^{bm}$. (Note that the final tensor is not an element of $\mathcal{T}_r{}^d{}_{mas}{}^{bm}$. There is no such set.)

4. Index Substitution. Given any tensor, together with a choice of an index of that tensor and a choice of a Latin letter which does not appear as an index of that tensor, there is associated another tensor, an element of the set denoted by "$\mathcal{T}$" with all the indices of the original tensor, except that the chosen index letter is replaced by the chosen letter. This operation is indicated by simply changing the appropriate index letter on the tensor. Thus, for example, if we choose the subscript "c" of $\eta^a{}_{cbn}$ and the Latin letter "r", the result of index substitution is $\eta^a{}_{rbn}$, an element of $\mathcal{T}^a{}_{rbn}$.

These four are the only (algebraic) operations available on tensors. What remains is to write the list of properties these operations satisfy.

1. *Conditions on addition.* Addition is associative and commutative. Within each set there is an additive identity (written "0", with indices suppressed), and each tensor has an additive inverse (written with a "−"). In

short, each of our sets is an *abelian* group under addition.

$$(\alpha^a{}_b + \beta^a{}_b) + \gamma^a{}_b = \alpha^a{}_b + (\beta^a{}_b + \gamma^a{}_b)$$

$$\alpha^a{}_b + \beta^a{}_b = \beta^a{}_b + \alpha^a{}_b$$

$$\alpha^a{}_b + 0 = \alpha^a{}_b$$

$$\alpha^a{}_b + (-\alpha^a{}_b) = \alpha^a{}_b - \alpha^a{}_b = 0$$

2. *Conditions on outer product.* Outer product is associative and distributive over addition.

$$\mu^{ab}(\nu_c{}^m\tau_r) = (\mu^{ab}\nu_c{}^m)\tau_r$$

$$\mu^{ab}(\alpha_n{}^r + \beta_n{}^r) = \mu^{ab}\alpha_n{}^r + \mu^{ab}\beta_n{}^r$$

$$(\sigma_n{}^r + \tau_n{}^r)\gamma^{ab} = \sigma_n{}^r\gamma^{ab} + \tau_n{}^r\gamma^{ab}$$

3. *Conditions on contraction.* The operation of contraction commutes with all the operations. Thus, if $\alpha^a{}_b + \beta^a{}_b = \gamma^a{}_b$ then $\alpha^b{}_b + \beta^b{}_b = \gamma^b{}_b$. (On the right is the sum of two elements of $\mathcal{T}$; on the left, contraction applied to an element of $\mathcal{T}^a{}_b$.) If $\mu_{ab}{}^{crs} = \sigma_{ab}{}^c\tau^{rs}$, then $\mu_{ab}{}^{ars} = \sigma_{ab}{}^a\tau^{rs}$. If $\alpha^a{}_b = \kappa^a{}_b{}^d{}_d$ and $\beta^c{}_d = \kappa^b{}_b{}^c{}_d$, then $\alpha^b{}_b = \beta^d{}_d$. (This double contraction would be written $\kappa^b{}_b{}^d{}_d$.) The result of substituting "r" for "b" in $\lambda^c{}_{bc}$ is the same as the result of contracting $\lambda^a{}_{rc}$ over "a" and "c".

4. *Condition on index substitution.* The operation of index substitution commutes with all the operations. The result of substituting one index for another, followed by substituting the other for the one, is the original tensor.

These conditions are easy to remember because they are suggested by the notation.

5. The set of scalars, $\mathcal{T}$, is the set of real numbers, and addition and outer product for scalars is addition and multiplication of real numbers.

   It follows immediately that *each of our sets is actually a vector space (over the real numbers).* Scalar multiplication is outer product by elements of $\mathcal{T}$. Furthermore, condition 4 implies that, *e.g.*, $\mathcal{T}^r{}_{ac}$ and $\mathcal{T}^s{}_{bn}$ are isomorphic vector spaces (the isomorphism obtained by index substitution). Furthermore, conditions 2 and 3 imply that the operation of contraction and outer product are linear mappings on the appropriate vector spaces.

6. Every tensor can be written as a (never unique) sum of outer products of vectors.

   On the outer product $\mu^a{}_{cdr}{}^s\alpha_u{}^v$, contract "s" with "u", and "v" with "r", to obtain an element of $\mathcal{T}^a{}_{cd}$ which may be written $\mu^a{}_{cdr}{}^s\alpha_s{}^r$. Thus, every element of $\mathcal{T}^a{}_{cdr}{}^s$ defines a linear mapping from the vector space $\mathcal{T}_s{}^r$ to the vector space $\mathcal{T}^a{}_{cd}$.

7. $\mathcal{T}^a{}_{cdr}{}^s$ consists precisely of the set of linear mappings from $\mathcal{T}_s{}^r$ to $\mathcal{T}^a{}_{cd}$, and similarly for other index combinations. Thus, for example, $\mathcal{T}_a$ is the dual of $\mathcal{T}^a$ (*i.e.*, the set of linear mappings from $\mathcal{T}$ to the reals). Further, $\mathcal{T}^a{}_b$ is the set of linear mappings from $\mathcal{T}^b$ to $\mathcal{T}^a$ (which, since $\mathcal{T}^b$ is isomorphic to $\mathcal{T}^a$, is the same as the set of linear mappings from $\mathcal{T}^b$ to $\mathcal{T}^b$). From this point of view, $\alpha^a{}_b\beta^b{}_c$ is the composition of a linear-mapping-from-$\mathcal{T}^c$-to-$\mathcal{T}^b$ with a linear-mapping-from-$\mathcal{T}^b$-to-$\mathcal{T}^a$ to obtain a linear-mapping-from-$\mathcal{T}^c$-to-$\mathcal{T}^a$. Further, $\alpha^a{}_a$ is the trace operation on linear mappings.

A **tensor algebra** is a collection of sets, on which four operations are defined, subject to seven conditions, as described above. We remark that the conditions are redundant. We have proceeded in this way in order to put the basic algebraic facts about tensors together on an equal footing.

The following central result summarizes much of linear algebra:

Theorem. Given any finite dimensional vector space V, there exists a tensor algebra with its $\mathcal{T}^a$ isomorphic to V. Furthermore, an isomorphism between the "$\mathcal{T}^a$'s" of two tensor algebras extends uniquely to an isomorphism of the tensor algebras.

In intuitive terms, a tensor algebra produces a framework for describing everything which can be obtained, starting from a single vector space, using linear mappings and tensor products. It equates naturally isomorphic things. It describes neatly the algebraic manipulations available on these objects.

Why didn't we require that outer product be commutative? Because it isn't defined. For example, $\alpha^{ab}\beta_c{}^d$ and $\beta_c{}^d\alpha^{ab}$ lie in different vector spaces, so equality is meaningless. We can, however, recover commutativity of outer product by introducing an additional convention. Let, for example, $\tau^a{}_{bc}$ be a tensor, and, by condition 6, write:

$$\tau^a{}_{bc} = \alpha^a\beta_b\gamma_c + \cdots + \mu^a\nu_b\kappa_c.$$

Then we can associate with this $\tau^a{}_{bc}$ an element, $\gamma_c\alpha^a\beta_b + \cdots + \kappa_c\mu^a\nu_b$, of $\mathcal{T}_c{}^a{}_b$. Thus, we obtain an isomorphism between $\mathcal{T}^a{}_{bc}$ and $\mathcal{T}_c{}^a{}_b$. Similarly, we have an isomorphism between any "$\mathcal{T}$" and any other obtained by changing the order of the indices, preserving their subscript or superscript status. We make use of these isomorphisms as follows. We permit ourselves to write equality of elements of different vector spaces, isomorphic as above, if the elements correspond under the isomorphism above. By this extended use of equality, for example, $\mu^a\nu_b = \nu_b\mu^a$. Similarly, we allow ourselves to add elements of isomorphic (as above) vector spaces by adding within one of the two spaces. Thus, $\gamma^a{}_{bc} = \delta_b{}^a{}_c + \sigma_{cb}{}^a$ means "carry $\sigma_{cb}{}^a$ to $\mathcal{T}_b{}^a{}_c$ via the isomorphism, and there add to $\delta_b{}^a{}_c$". The result, carried to $\mathcal{T}^a{}_{bc}$ via the isomorphism, is equal to $\gamma^a{}_{bc}$.

As a final example, let ξ_{ab} be a tensor. Use index substitution to successively obtain ξ_{ac}, ξ_{dc}, ξ_{da}, and ξ_{ba}. If it should happen that $\xi_{ab} = \xi_{ba}$, then our tensor is called **symmetric**. If it should happen that $\xi_{ab} = -\xi_{ba}$, then our tensor is called **antisymmetric**. In general, such a tensor will be neither symmetric nor antisymmetric. It can always be written uniquely as the sum of a symmetric and an antisymmetric tensor: $\xi_{ab} = \frac{1}{2}(\xi_{ab} + \xi_{ba}) + \frac{1}{2}(\xi_{ab} - \xi_{ba})$. These facts are familiar from matrix algebra.

3. Tensor fields

Our next task is to combine our discussion of manifolds with that of tensor analysis.

Let M be a manifold. A **scalar field** on M is simply a real-valued function on M. That is, a scalar field assigns a real number to each point of M. It turns out, however, that arbitrary scalar fields aren't very interesting: what is interesting are things we shall call *smooth* scalar fields. Let α be a scalar field on M. To define smoothness of α, we must make use of the charts on M. Thus, let (U, ψ) be a chart. Then the points of M in U are labeled, by this chart, by coordinates $(x^1, \ldots, x^n)$. Since α is just a function on M (and hence, also a function on U), we may represent α as one real function of the n variables $(x^1, \ldots, x^n)$. Formally, this function is $\alpha \circ \psi^{-1}$. We say that the scalar field is **smooth** if, for every chart (in the definition of the manifold M), $\alpha \circ \psi^{-1}$ (a function of n variables) is C^∞ (all partials of all orders exist and are continuous). We denote by $\mathcal{S}$ the collection of smooth scalar fields on M. It is obvious that sums and products of smooth scalar fields yield smooth scalar fields. It is also clear that there are many nonconstant smooth scalar fields on M.

We wish next to define the notion of a *vector on a manifold*. It is convenient, for motivation, to first recall some things about vectors in Euclidean space. A vector in Euclidean space can be represented by its components, $(\xi^1, \ldots, \xi^n)$. This representation is not, however, very convenient for a discussion of vectors on a manifold, because of the enormous freedom available on the choice of charts (with respect to which components could be taken) on a manifold. We would like, therefore, to think of some description of vectors in Euclidean space which refers less explicitly to components. Let $\alpha(x^1, \ldots, x^n)$ be a smooth function of n variables, *i.e.*, a smooth function on our Euclidean space. Then we can consider the directional-derivative of this function in the direction of our vector, evaluated at the origin:

$$\xi(\alpha) = \left(\xi^1 \frac{\partial \alpha}{\partial x^1} + \cdots + \xi^n \frac{\partial \alpha}{\partial x^n} \right|_{x=0} \tag{1}$$

Thus, given a vector in Euclidean space, $\xi(\alpha)$ is a number for each function α.

It is immediate from the properties of partial-derivatives that $\xi(\alpha)$, regarded as a mapping from smooth functions on Euclidean space to the reals, satisfies the following conditions:

1. $\xi(\alpha + \beta) = \xi(\alpha) + \xi(\beta)$.
2. $\xi(\alpha\beta) = \alpha|_{x=0}\, \xi(\beta) + \beta|_{x=0}\, \xi(\alpha)$.
3. If $\alpha = \text{const}$, $\xi(\alpha) = 0$.

Thus, a vector in Euclidean space defines a mapping (namely, the directional derivative) from smooth functions on Euclidean space to the reals, satisfying the three conditions above. We now establish a sort of converse, to the general effect that these three conditions characterize vectors in Euclidean space. More precisely, we claim that:

A mapping from smooth functions on Euclidean space to the reals satisfying the three conditions above is of the form (1) for some $(\xi^1, \ldots, \xi^n)$.

Proof: Suppose we have such a mapping. Define n numbers by $\xi^1 = \xi(x^1)$, $\ldots$, $\xi^n = \xi(x^n)$. We shall show that, for this choice, (1) is valid for any smooth function α. Write such an α in the form

$$\alpha = \underline{\alpha} + \sum_i \alpha_i x^i + \sum_{ij} \alpha_{ij}(x) x^i x^j \tag{2}$$

where $\underline{\alpha}$ is a constant, the α_i $(i = 1, \ldots, n)$ are constants, and α_{ij} $(i, j = 1, \ldots, n)$ are smooth. Then, using conditions 1 and 2 above,

$$\xi(\alpha) = \xi(\underline{\alpha}) + \sum_i \left[\xi(\alpha_i)x^i + \alpha_i\xi(x^i)\right] + \sum_{ij} \left[\xi(\alpha_{ij})x^ix^j + \alpha_{ij}\xi(x^j)x^i + \alpha_{ij}\xi(x^i)x^j\right].$$

Now use the condition 3 and evaluate at the origin:

$$\xi(\alpha) = \sum_i \alpha_i\xi(x^i) = \sum_i \alpha_i\xi^i$$

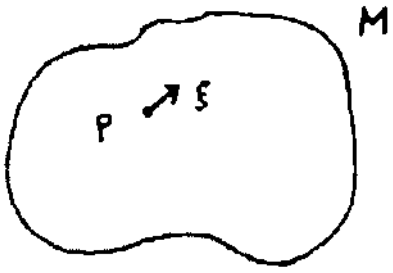
This is the left side of Eqn. (1). But, from (2), this is also clearly the right side of (1). Hence, (1) holds.

We have now completed our component-independent characterization of vectors in Euclidean space. This characterization is the motivation for the definition, which follows, of vectors on a manifold.

Fix a point p of our manifold M. A (**contravariant**) **vector** in M at p is a mapping $\xi : \mathcal{S} \to \mathbb{R}$ from the smooth scalar fields on M to the reals satisfying:

- $\xi(\alpha + \beta) = \xi(\alpha) + \xi(\beta)$.
- $\xi(\alpha\beta) = \alpha|_p\,\xi(\beta) + \beta|_p\,\xi(\alpha)$.
- If $\alpha = \text{const}$, then $\xi(\alpha) = 0$.

It is clear from the discussion above that the contravariant vectors at a point p of an n-dimensional manifold form an n-dimensional vector-space.



Intuitively, a contravariant vector at a point-of-a-manifold points in a "direction", just as do vectors in Euclidean space.

We now bring in the tensor algebras. At each point p of our manifold M, we have the n-dimensional vector-space of contravariant vectors at p. We now identify this vector space with the vector space $\mathcal{T}^a$ of a tensor algebra ("identify" means "make isomorphic to"). Then, as we have seen, we acquire a unique tensor algebra. These objects will be called *tensors* (in the manifold M) at the point p. Thus, a "scalar at p" is just a real number; a "vector ξ^a at p" is a contravariant vector, *i.e.*, a mapping from smooth scalar fields on M to reals satisfying our three conditions; a "vector η_a at p" is (among other things) a linear mapping from $\mathcal{T}^a$ at p to the reals. One repeats these remarks for each point of our manifold M. Thus, we have a notion of "a tensor at a point of M".

The next thing we must do is extend the notion of "a tensor at a point" to that of a tensor field. A **tensor field** on M assigns, to each point of M, a tensor at that point, where the tensors at the various points of M all have the same index-structure. For example, let $\alpha_{ac}{}^b{}_d{}^q$ be a tensor field on M. Then we have assigned, to each point of M, a tensor $\alpha_{ac}{}^b{}_d{}^q(p)$, at p. In particular, a scalar field on M, by this definition, assigns a scalar (*i.e.*, a real number) to each point of M. That is to say, a scalar field on M is just a real-valued function on M, which coincides with our original definition.

We observe that the operations defined within a tensor algebra extend from "tensors at a point" to tensor fields. Thus, let ${\alpha_a}^b$ and ${\beta_a}^b$ be tensor fields on M. Their sum, written as ${\alpha_a}^b + {\beta_a}^b$, is a tensor field on M which assigns, to the point p of M, the sum of the tensors assigned to p by ${\alpha_a}^b$ and ${\beta_a}^b$. That is, $({\alpha_a}^b + {\beta_a}^b)(p) = {\alpha_a}^b(p) + {\beta_a}^b(p)$. Similarly, outer product, contraction, and index-substitution are well-defined operations on tensor fields. Furthermore, the various properties of tensors in a tensor algebra extend to properties of tensor fields. Thus, addition of tensor fields is associative and commutative. The zero tensor field (which assigns to each point the zero-tensor) is an additive identity. Outer products for tensor fields is associative and distributive. Contraction and index-substitution commute with everything in sight, and so on.

We saw a few pages ago that the notion of a scalar field is not very useful: what is useful is that of a *smooth* scalar field. This is perhaps not surprising, since the heart of a manifold is its smoothness structure, and one might expect that compatibility with that smoothness structure should be crucial. Since scalar fields were rejected in favor of smooth scalar fields, it is natural to ask whether tensor fields can somehow be rejected in favor of smooth tensor fields. This is indeed possible: we now do it.

Consider a contravariant vector field ξ^a. Let α be any smooth scalar field. Then, for each point p of our manifold, $\xi^a(p)$ is a contravariant vector at p. Hence, it assigns, to the smooth scalar field α a number. Keeping ξ^a and α fixed, and repeating at each point p, we assign a real number to each point of M. In other words, we obtain a scalar field on M. (This scalar field is, of course, just the directional derivative of α in the direction of the vector field ξ^a.) We say that our contravariant vector field on M can be regarded as a mapping from smooth scalar fields on M to scalar fields on M. We say that our contravariant vector field ξ^a on M is **smooth** if it assigns, to each smooth scalar field on M a *smooth*-scalar field. Thus, we have extended the notion of smoothness from scalar fields to contravariant vector fields. A vector field η_a will be called *smooth* if, for each smooth ξ^a (smoothness for these just defined above), $\eta_a\xi^a$ is smooth. Finally, a tensor field, *e.g.*, ${\kappa_{ac}}^r{}_s{}^q$, will be called *smooth* if ${\kappa_{ac}}^r{}_s{}^q\xi^a\eta^c\lambda_r\omega^s\rho_q$ is a smooth scalar field for any smooth vector fields ξ^a, η^c, λ_r, ω^s, ρ_q. Note the way that the notion of smoothness permeates the tensor fields, starting from the scalar fields. This sort of thing is common in differential-geometry. We now notice that all the tensor operations, applied to smooth tensor fields, yield smooth tensor fields.

By the last sentence of the previous paragraph, if one starts with only smooth tensor fields, one never gets anything but smooth fields. It turns out in practice that it is just the smooth fields which are useful in physics. It is convenient, for this reason, not to have to always carry around the adjective "smooth". We shall henceforth omit it. When we say "tensor field", we mean "*smooth* tensor field" unless otherwise specified.

As we shall see shortly, manifolds and tensor fields are remarkably well-suited for the description of the arena in which physics takes place, and of the physics itself, respectively.

4. Derivative Operators

What is so great about smoothness? What was gained by our restricting consideration to smooth tensor fields? In the case of scalar fields, what we gained by requiring smoothness was the ability to take (directional) derivatives. This observation (together with the connotation of the word "smooth") suggests that it should be possible—and perhaps useful—to introduce the concept of the derivative of a (smooth) tensor field.

We can think of an n-dimensional manifold as similar to an n-dimensional surface. Thus, there are n directions in which one can move from a given point. That is, given a tensor field, there are n directions in which one can take the derivative of that tensor field. These remarks suggest that the operation "taking a derivative" should increase the number of indices of a tensor field by one. There is a more explicit way of seeing this. Let α be a scalar field. What should "the derivative of α" look like (as regards index-structure)? It would be right, somehow, if the derivative of α had one lowered index, *e.g.*, if it looked something like $\nabla_a\alpha$, for then $\xi^a\nabla_a\alpha$, for any contravariant vector field ξ^a, would naturally represent the directional derivative of α by ξ^a. The index on the "derivative operator" ∇_a provides a place to park the index of ξ^a, *i.e.*, it reflects the freedom in the direction in which the derivative can be taken. Thus, it seems reasonable, as a first try, to try define the "derivative operator" as an operator, with one down index, which acts on tensor fields. Having made this observation, it would be difficult to wind up with any definition for a derivative operator other than that which follows.

A **derivative operator**, ∇_a, is a mapping from (smooth) tensor fields to (smooth) tensor fields, which adds one lowered index, and which satisfies the following conditions:

1. The derivatives of the sum of two tensor fields is the sum of the derivatives. For example,

$$\nabla_a(\alpha^r{}_{cq} + \beta^r{}_{cq}) = \nabla_a(\alpha^r{}_{cq}) + \nabla_a(\beta^r{}_{cq}).$$

2. The derivative of the outer-product satisfies the Leibniz rule. For example,

$$\nabla_a(\mu^u{}_c{}^d\nu_{qv}) = \mu^u{}_c{}^d\nabla_a\nu_{qv} + \nu_{qv}\nabla_a\mu^u{}_c{}^d.$$

3. The derivative operation commutes with contraction and index substitution. For example, if $\beta = \alpha^a{}_a$ and $\gamma_c{}^a{}_b = \nabla_c\alpha^a{}_b$, then $\nabla_c\beta = \gamma_c{}^a{}_a$, and similarly for index substitution.

4. For any scalar field α and contravariant vector field ξ^a the scalar field $\xi^a\nabla_a\alpha$ is just the directional derivative.

5. Two derivative operations, applied to a scalar field, commute. That is, $\nabla_a\nabla_b\alpha = \nabla_b\nabla_a\alpha$. (Derivatives will <u>not</u> commute, in general, applied to tensor fields with more indices.)

Except possibly for the last, these are all natural-looking "derivative-like" conditions. It is easy to define things. What is usually more difficult is to show that these things exist, and to discover to what extent they are unique. We now ask these questions for derivative operators.

Every manifold one is ever likely to meet possesses a derivative operator. (Precisely, a manifold possesses at least one derivative operator if and only if it is paracompact.) In fact, it is a nontrivial job just to construct an example of a manifold which does not have any derivative operators. We shall henceforth deal only with manifolds which have at least one derivative operator. (In fact, this restriction will not be needed in any formal sense, but only to make discussions easier. We shall use the assumed existence of a derivative operator

in three cases: to define Lie derivatives, exterior derivatives, and the derivative operator defined by a metric. Lie and exterior derivatives exist whether or not derivative operators exist. That the existence of a metric implies the existence of a derivative operator follows by a comparatively simple argument.)

A more interesting question is this: *how unique are derivative operators?* It turns out that derivative operators are never unique, but that the non-uniqueness can be expressed in a very simple way. We now proceed to explain this remark in more detail.

Suppose we had two derivative operators on the manifold M, ∇_a and ∇'_a. How are they related? We first consider the difference between their actions on a scalar field: $\nabla'_a\alpha - \nabla_a\alpha$. To draw conclusions about this expression, we contract it with an arbitrary contravariant vector field ξ^a, to obtain $\xi^a\nabla'_a\alpha - \xi^a\nabla_a\alpha$. This is the difference between two scalar fields. But, by condition 4, in the definition of a derivative operator, these two scalar fields are the same. So, the difference is zero. That is to say, $\xi^a(\nabla'_a\alpha - \nabla_a\alpha)$ vanishes for any ξ^a. Hence, $\nabla'_a\alpha - \nabla_a\alpha = 0$. That is, $\nabla'_a\alpha = \nabla_a\alpha$. In words, any two derivative operators have, by condition 4, the same action on scalar fields.

We next compare the actions of our two derivative operators on contravariant vector fields. Consider $\nabla'_a\xi^b - \nabla_a\xi^b$. We cannot conclude, as above, that this vanishes. However, it is true that this expression is linear in ξ^b. This follows from the following calculation

$$\begin{aligned}
&(\nabla'_a - \nabla_a)(\alpha\xi^b + \eta^b) \\
&\quad = \quad \nabla'_a(\alpha\xi^b) - \nabla_a(\alpha\xi^b) + \nabla'_a(\eta^b) - \nabla_a(\eta^b) \\
&\quad = \quad \xi^b\nabla'_a(\alpha) + \alpha\nabla'_a(\xi^b) - \xi^b\nabla_a(\alpha) - \alpha\nabla_a(\xi^b) + (\nabla'_a - \nabla_a)\eta^b \\
&\quad = \quad \alpha(\nabla'_a - \nabla_a)\xi^b + (\nabla'_a - \nabla_a)\eta^b
\end{aligned}$$

where, in the last step, we have used the fact that two derivative operators coincide on scalar fields. Hence, there is some tensor field ${C^m}_{ab}$ such that

$$\nabla'_a\xi^b - \nabla_a\xi^b = -{C^b}_{ac}\xi^c \tag{3}$$

for all ξ^a. In short, derivative operators needn't be the same on contravariant vector fields, but their difference is expressible simply in terms of a certain tensor field.

Next, vector fields with one down index (covariant vector fields). Contract $\nabla'_a\eta_b - \nabla_a\eta_b$ with an arbitrary ξ^b, and use the properties of derivative operators:

$$\begin{aligned}
\xi^b(\nabla'_a\eta_b - \nabla_a\eta_b) \quad &= \quad \nabla'_a(\eta_b\xi^b) - \eta_b\nabla'_a\xi^b - \nabla_a\xi^b\eta_b + \eta_b\nabla_a\xi^b \\
&= \quad -\eta_b(\nabla'_a\xi^b - \nabla_a\xi^b) \\
&= \quad -\eta_b(-{C^b}_{ac}\xi^c) \\
&= \quad +\eta_m{C^m}_{ab}\xi^b.
\end{aligned}$$

Since ξ^b is arbitrary, we have

$$\nabla'_a\xi^b - \nabla_a\xi^b = \eta_m{C^m}_{ab} \tag{4}$$

Thus, the actions of our derivative operators on covariant vector fields also differ in a simple way, and, furthermore, the same tensor ${C^m}_{ab}$ as for contravariant vector fields comes into play.

We next derive, from the fifth property of derivative operators, a certain property of this ${C^m}_{ab}$. Let α be any scalar field. Then

$$\begin{aligned}
\nabla'_a\nabla'_b\alpha - \nabla_a\nabla_b\alpha \quad &= \quad \nabla'_a\nabla_b\alpha - \nabla_a\nabla_b\alpha \\
&= \quad (\nabla'_a - \nabla_a)\nabla_b\alpha = {C^m}_{ab}\nabla_m\alpha
\end{aligned}$$

where, in the first step, we have used the fact that all derivative operators coincide on scalars. Now, the left side of this equation remains invariant if "a"

and "b" are switched. So, the right side must. But this holds for all scalar fields α, and so

$$C^m{}_{ab} = C^m{}_{ba} \tag{5}$$

What remains is to extend (3) and (4) to general tensor fields. This is done as follows. We wish to evaluate, say, $(\nabla'_a - \nabla_a)\alpha_{cd}{}^{rs}{}_u$. We contract it with vectors $\xi^c\eta^d\mu_r\nu_s\lambda^u$. We now differentiate by parts, just as we did in the calculation preceding Eqn. (4). The result is to shift the derivative operators to the vectors. Now use (3) and (4) to get rid of the derivative operators in favor of the $C^m{}_{ab}$. Finally, use the fact that the vectors are arbitrary—so they can be removed. The result (which, with a little thought, can be seen without actually doing the calculation) is

$$\begin{aligned}(\nabla'_a - \nabla_a)\alpha_{cd}{}^{rs}{}_u \;\; = \;\; & C^m{}_{ac}\alpha_{md}{}^{rs}{}_u + C^m{}_{ad}\alpha_{cm}{}^{rs}{}_u \\ & -C^r{}_{am}\alpha_{cd}{}^{ms}{}_u - C^s{}_{am}\alpha_{cd}{}^{rm}{}_u + C^m{}_{au}\alpha_{cd}{}^{rs}{}_m \end{aligned} \tag{6}$$

Note the form of the right hand of Eqn. (6). There appears a term corresponding to each index of $\alpha_{cd}{}^{rs}{}_u$. Raised indices are treated just like ξ^b is in (3), while lowered indices are treated just like η_b is in (4). Thus, Eqns. (3) and (4), which are now special cases of (6), also make it easy to remember (6). Note also that the fact that all derivative operators coincide on scalar fields is a special case of Eqn. (6): no indices leads to no terms on the right.

We conclude that, given two derivative operators, ∇_a and ∇'_a, there is a tensor field $C^m{}_{ab}$ satisfying (5) such that the actions of these derivative operators are related by Eqn. (6). A converse is immediate. Suppose we have a derivative operator ∇_a, and suppose we choose any (smooth) tensor field $C^m{}_{ab}$ satisfying Eqn. (5). Then we can define a new operation, ∇'_a, on tensor fields by Eqn. (6). (for (6) can be interpreted now as giving the action of ∇'_a on any tensor field). It is immediate that this ∇'_a, so defined, satisfies the five conditions for a derivative operator. Thus, a derivative operator together with a $C^m{}_{ab}$ satisfying (5) defines another derivative operator.

In short, derivative operators are never unique, but we have good control over the non-uniqueness.

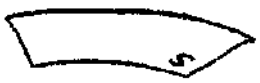
Part II.
Mechanics

5. Configuration Space

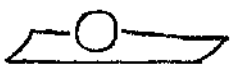
The things in the physical world in which we shall be interested will be called *systems*. It is difficult, and probably futile, to try to formulate with any precision what a system is. Roughly speaking, a system is a collection of things (*e.g.*, objects, particles, fields) under study which, in some sense, do not interact with the rest of the universe. This splitting of the universe into separate systems one singles out for individual study is certainly arbitrary to some extent—that is to say, it is probably subject to what is fashionable in physics, and the mode of splitting probably changes as physics evolves. Nonetheless, we have to study something, and the things we study we call systems.

Examples of systems:

1. A free, point particle in Euclidean space.
2. A three-dimensional harmonic oscillator. That is, a point particle in Euclidean space, subject to a force directed towards the origin, and having strength proportional to the distance of the particle from the origin.
3. A point particle constrained to lie on a smooth two-dimensional surface S in Euclidean space.



4. A point particle in Euclidean space for which, for some reason, is unable to acquire any velocity component in the z-direction.
5. A ball free to roll (but not slide) over a horizontal plane in Euclidean space.



6. Electric and magnetic fields (say, without sources) in Euclidean space.

The systems listed above are not very exotic, and not very complicated. Neither are they described precisely.

The next step in the description of a system (after deciding what the system is) is the assignment to our system of a *configuration space*. The **configuration**

space, C, is supposed to be a manifold, the points of which represent "configurations" of the system. There appears to be no definite rule, given an actual physical system (*e.g.*, a Swiss watch), for deciding what its space of configurations, C, should be. Apparently, this is essentially an art. For a simple mechanical system, the configurations are normally distinguished by what can be seen in an instantaneous photograph. Thus, in examples (1) and (2) above, one would normally choose for the set C the set of "locations of the particle", *i.e.*, the set of points in Euclidean 3-space. In example (3), one could, I suppose, also choose for C the set of points of Euclidean space. But that isn't a particularly good choice. One should instead, choose the "locations actually available to the particle", *i.e.*, the set of points on the surface S. It is not uncommon, in examples such as (3) above, to first make the "wrong" choice for configuration space, *i.e.*, to first choose Euclidean space. One then corrects oneself: he realizes that only certain of these configurations, namely, those on S, are actually available to the particle. This process—first choosing "too large" a configuration space, and then imposing restrictions—is called that of introducing *holonomic constraints*. It is my view that much is gained—and nothing lost—by being careful in the beginning. We shall be careful and, thus, we shall be able to avoid what are called holonomic constraints.

Configuration space is, as stated above, to be more than just a set of points—it is to be a manifold. That is to say, the set of configurations must be endowed with a smoothness structure, *i.e.*, with a collection of charts satisfying our four conditions for a manifold. Here, again, there appears to be no prescription, given a physical system and having decided what its set of configurations is to be, to determine the manifold structure. There is genuine content in this choice of smoothness structure: it, for example, tells us when two configuration are to be regarded as "close". In examples (1) and (2) above, one would naturally choose for the manifold structure on C the natural structure on Euclidean space. That is, in these examples, one would choose for C the manifold $\mathbb{R}^3$. In example (3), one would select charts appropriate to the embedding of S in Euclidean space. That is, in this example, C would be a 2-dimensional manifold (*e.g.*, a sphere, torus, plane, *etc.*, depending on what S is).

Note that the first two examples are essentially different physical systems, although they have the same configuration space.

Example (4) provides an illustration of these issues. Let us first choose for the set C the set of points in Euclidean space. What should we choose for the manifold structure? On the one hand, we could just choose the natural manifold structure on Euclidean space. Then C would be $\mathbb{R}^3$. There is, however, another reasonable option. A particle in this example remains in the same $z =$ constant plane all its life. Perhaps, then, we should regard each such plane as a separate "piece" of configuration space. Thus, perhaps we should endow C with the following manifold structure: C is the (disjoint) union of an infinite stack of 2-planes (each labeled by its z value). For this choice, C would be two-dimensional. For these two choices, we have the same set C, but different smoothness structures. There is even a third point of view. One could claim that the system hasn't been specified with enough care: that one should also be told the z-plane in which the particle beings. This additional information is to be regarded as part of the statement of what the system is. Then, of course, one would choose for C that one plane, *i.e.*, C would be $\mathbb{R}^2$ (since the only configurations available to the particle need be included in C). Thus, one could reasonably argue for any of the three configuration spaces for example (4): $\mathbb{R}^3$, the disjoint union of an uncountable collection of $\mathbb{R}^2$'s, or a simple $\mathbb{R}^2$. My taste leans mildly toward the second of these three choices.

To specify the "configuration of the system" in example (5), one must state where (on the plane) the sphere is, and the orientation of the sphere in space. "Where the plane is" would be represented by a point on the manifold $\mathbb{R}^2$, and the orientation (which could be described by the rotation necessary to obtain

the desired orientation from some fixed orientation) by a point on the manifold O^3 (pp 3). Thus, a natural choice for the configuration space in this example would be the product of manifolds $\mathbb{R}^2 \times O^3$. We shall shortly define precisely the product of manifolds.)

Example (6) is a bit tricky. We are dealing, here, with vector fields $\vec{E}$ and $\vec{B}$ in Euclidean space, both of which have vanishing divergence. Here, the difficulty in giving a prescription for selecting configuration space becomes more clear. Perhaps the best one can do is try various possibilities for configuration space, and see which one works best. The one which has been found to work best is this: choose for C the set of all vector fields $\vec{B}$ in Euclidean space which have vanishing divergence. (Holonomic constraint types would say this: choose first all vector fields $\vec{B}$, and then restrict to the ones which have vanishing divergence.) This is the set C; what is the smoothness structure? It turns out that the only reasonable choice makes C what is called an *infinite-dimensional manifold*. We will not deal here with infinite-dimensional manifolds (because of technical complications). Hence, we shall not be able to deal in detail with example (6).

We summarize: reasonable people could disagree, given a physical system, on what configuration space is appropriate. For most systems, however, one choice is particularly natural, and so we shall allow to speak of the configuration space C of a system.

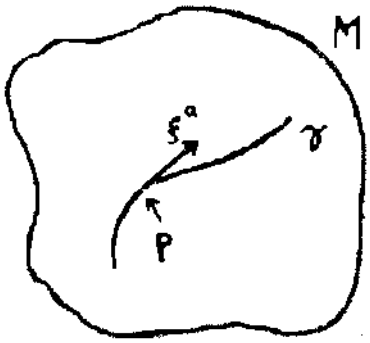
The dimension of the configuration space of a system is called the **number of degrees of freedom** of the system.

As times marches on, our system evolves, *i.e.*, it passes from one instant of time to the next, through various configurations. The mathematical description of this situation is via the notion of a curve. A **curve** on a manifold M is a mapping $\gamma : \mathbb{R} \to M$, where $\mathbb{R}$ is the reals. That is, for each real number t, $\gamma(t)$ is a point of M. So, as t sweeps through $\mathbb{R}$, $\gamma(t)$ describes our curve. A notion more useful is that of smooth curves, where smoothness for curves is defined in terms of smooth scalar fields. Let α be a smooth scalar field, so $\alpha : M \to R$. Then α. Then $\alpha \circ \gamma$ is a mapping from $\mathbb{R}$ to $\mathbb{R}$, intuitively, "the function α evaluated, in terms of t, along the curve". We say that the curve γ is **smooth** if, for every smooth scalar field α, $\alpha \circ \gamma$ (one real function of one variable) is C^∞. Thus, the evolution of our system is described by a smooth curve γ, with parameter T representing time, on the configuration space C. Thus, whereas examples (1) and (2) have the same configuration space, the actual curves giving the evolution are different in the two examples.

We now need some more mathematics: the notion of the *tangent vector to a curve*. Let γ be a curve on M. Fix a number t_0, and set $p = \gamma(t_0)$. Now let α be a scalar field on M, and consider the number

$$\frac{d}{dt}\alpha(\gamma(t))\bigg|_{t_0} .$$

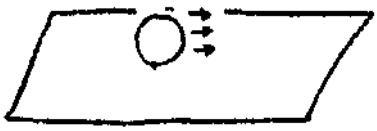
Geometrically, this is the rate of change of the function α, with respect to t, along the curve, evaluated at t_0. It is immediate from the properties of the derivative (of functions of one variable) that this mapping from scalar fields on M to the reals satisfies the three conditions on page 8. Hence, this mapping defines a contravariant vector γ^a at p. This vector is called the **tangent vector** to the curve γ at the point $\gamma(t_0)$. Geometrically, the tangent vector is tangent to the curve, and is large when the curve "covers a lot of M for each increment of t".



We now return to our system, its configuration space C, and some curve γ giving the evolution of the system. The tangent vector to this curve at $\gamma(t_0)$ is called the **velocity** of the system at time t_0, written v^a. This just generalizes the notion of the velocity, *e.g.*, of a particle moving in Euclidean space. Note that an evolving system, by our description, defines, at each moment of time, a configuration and a velocity.

As an illustration of the notion of velocity, we describe the set-up for *Lagrangian mechanics*. Let C be the configuration space for a system. The set of all pairs (q, v^a), where q is a point of C and v^a is a contravariant vector at q is called the **tangent bundle** of C. (This is technically not quite correct. The tangent bundle of a manifold is actually another manifold, having the dimension twice that of the original manifold. The set of pairs (q, v^a) above is actually just the underlying point-set of the tangent bundle.) The starting point for Lagrangian mechanics is the introduction of a certain function $L(q, v^a)$ on the tangent bundle of configuration space. This real-valued function is called the **Lagrangian** of the system. One sometimes writes the Lagrangian as $L(x, \partial x)$.

Are all points of the configuration space accessible to the system? In essentially all cases the answer is yes. In fact, that the answer be yes is one good criterion to use in selecting configuration space. Are all points of the tangent bundle accessible to the system? (More precisely, is it true that, given any point (q, v^a) of the tangent bundle, there exists a curve in configuration space, describing a possible evolution of the system, such that the curve passes through q and there has tangent vector v^a?) For many examples, the answer is yes, but it could be no.



Consider example (5) on page 13. Consider any point of configuration space, and a contravariant vector which represents a translation of the sphere, without permitting it to roll. Here is a point of the tangent bundle which is not allowed to the system, because the prescribed contravariant vector would represent a velocity inconsistent with the requirement that the ball not slide on the plane. When it is the case that not every point of the tangent bundle is accessible to the system, one says that "the system exhibits a constraint". (One sometimes calls this a *nonholonomic constraint*. For us, the adjective is redundant, so we

will not use it.) Thus, for example, for our three choices of the configuration space for example (4), the first has a constraint and the other two do not. There is certainly a sense in which the phrase "has a constraint" represents a property of the system itself more than a property of our description of the system. It appears to be difficult to make this remark precise. Does the point of the tangent bundle occupied by the system at one instant of time determine uniquely the future evolution of the system? (More precisely, do two curves, describing a possible evolution of the system, which pass through the same point of C for a given time t, and which there have the same tangent vector, necessarily coincide?) Again, examples suggest an answer yes, and this might be taken as a criterion for a good choice of configuration space. Again, this property is a rather complicated mixture of an aspect of Nature and an aspect we impose on Nature by our mode of description.

Suppose we have two systems, S_1 and S_2, and that we decide to regard them as a single system? (Note that nothing has happened physically: we just look at things differently.) If C_1 and C_2 are the configuration spaces of our systems, what should we choose for the configuration space of the combined system? The answer is $C = C_1 \times C_2$, a *product* of manifolds. We now define this. As a set, C consists of all pairs (p_1, p_2), where p_1 is a point of C_1 and p_2 is a point of C_2. We introduce charts. Let (U_1, ψ_1) be a chart on C_1, and (U_2, ψ_2) a chart on C_2. Consider the chart on C with $U = U_1 \times U_2$, and, for (p_1, p_2) in U, $\psi(p_1, p_2) = (\psi_1(p_1), \psi_2(p_2))$. (This last expression is an $(n_1 + n_2)$-tuple, where n_1 and n_2 are the dimensions of C_1 and C_2, respectively.) The collection of all charts of C compatible with these makes C a manifold. The dimension of C is, of course, $n_1 + n_2$. Thus, for example, the product of the manifold $\mathbb{R}^1$ and the manifold $\mathbb{R}^1$ is the manifold $\mathbb{R}^2$; The product of $\mathbb{R}^1$ and S^1 is the cylinder; The product of S^1 and S^1 is the torus. Thus, the combination of systems is represented, within our description, by the product of configuration space manifolds.

6. The Cotangent Bundle

It is convenient to interrupt briefly our study of mechanics at this point to introduce a little mathematics.

Let M be an n-dimensional manifold. We define, from M, a new $2n$-dimensional manifold Γ_M called the **cotangent bundle** of M. As a point-set, Γ_M consists of all pairs (q, p_a), where q is a point of M and p_a is a covariant vector in M at the point q. (It is not difficult to see, already at this point, why Γ_M will turn out to be $2n$-dimensional. It takes n dimensions to "locate" q in M, and, having chosen q, n more dimensions to locate p_a. (Recall that the vector space of covariant vectors at a point of an n-dimensional manifold is n dimensional.)) What remains is to introduce a collection of charts on this set Γ_M, and to verify that this collection of charts satisfies our four conditions for a manifold. Let (U, ψ) be any chart on M, so U is a subset of M and ψ, a mapping from U to $\mathbb{R}^n$. We associate, with this chart, a certain chart (U', ψ') (so U' will be a subset of Γ_M and ψ' a mapping from U' to $\mathbb{R}^{2n}$) on the set Γ_M. For U' we choose the collection of all pairs (q, p_a) with q in U. Then, for (q, p_a) in U', we set $\psi'(q, p_a) = (\psi(q), \kappa_1, \ldots, \kappa_n)$ where $\kappa_1, \ldots, \kappa_n$ are the n numbers such that

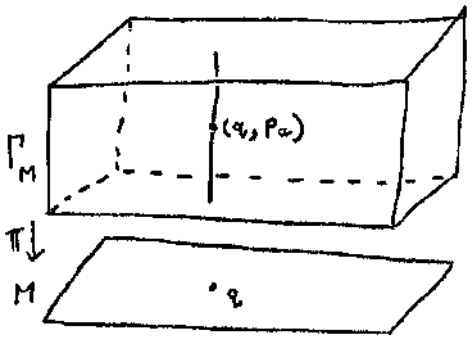
$$p_a = \kappa_1 \nabla_a x^1 + \cdots + \kappa_n \nabla_a x^n$$

where the right hand side is evaluated at q.

We now observe that these charts on Γ_M are all compatible with each other, and that they cover Γ_M. The collection of all charts on Γ_M compatible with these makes Γ_M a manifold, of dimension $2n$.

We repeat this construction in words. A chart on M labels points of M by n-tuples of real numbers. A point of Γ_M is a pair (q, p_a), where p_a is a covariant vector at the point q of M. Half the information needed to specify a point of Γ_M, namely q, is already labeled by n numbers, namely, the n numbers $\psi(q)$. For the other half, namely p_a, we observe that the gradients of the n coordinate functions define, at each point of U (in M), n covariant vectors which span the space of covariant vectors at that point. Hence, we can label p_a by the n numbers giving the expression for p_a, as a linear combination of the gradients of the coordinate functions. In this way, we obtain suitable charts on M.

There is a natural mapping, which we write π, from Γ_M to M. It is defined by $\pi(q, p_a) = q$. In words, π is the mapping which "forgets" what the covariant vector p_a is, but remembers the point q of M. A picture of this mapping, is that on the right. Each point q of M is located below a vertical line in Γ_M which represents all points of the form (q, p_a). Thus, in terms of this picture, π is the mapping which sends each point of Γ_M straight down to a point of M. It is immediate, for example, that π is always onto, and that it is one-to-one if and only if M is zero-dimensional.



This π is our first example of a mapping from one manifold to another. It turns out that some such mappings are more interesting than others. The interesting ones are the ones we shall call *smooth*. Our next task is to define

smoothness of a mapping from one manifold to another. Let $\tau : N \to M$ be a mapping from the manifold N to the manifold M. We say that this mapping is **smooth** if, for any smooth scalar field α on M, $\alpha \circ \tau$ (necessarily a real-valued function on N) is smooth. Note that $\alpha \circ \tau$ is the function on N whose value at a point x of N is the number assigned by α to the point $\tau(x)$ of M. One always defines smoothness of things on manifolds by asking that, when one does what one can with those things on smooth scalar fields, the result is smooth.

As an example of this notion of smoothness of mappings, we show that $\pi : \Gamma_M \to M$ is smooth. Let α be a smooth scalar field on M. Then $\beta = \alpha \circ \pi$ is the following function on Γ_M: $\beta(q, p_a) = \alpha(q)$. We must show that this function β is smooth. Let (U, ψ) be a chart on M, and let (U', ψ') be the corresponding chart on Γ_M. Then, in terms of the coordinates given by this chart, β is the following real function of $2n$ variables: $\beta(x^1, \ldots, x^n, \kappa_1, \ldots, \kappa_n) = \alpha(x^1, \ldots, x^n)$. Since α is smooth, the function (of n variables) on the right is smooth. Hence, the function (of $2n$ variables) representing β is smooth. Hence, β is a smooth scalar field on Γ_M. Hence, π is a smooth mapping.

The beautiful thing about the cotangent bundle of any manifold is that it possesses a natural covariant vector field. Since we are using Latin indices to label tensors on M, we shall, to avoid confusion, use Greek indices for tensors on Γ_M. This covariant vector field will be denoted A_α. Our next task is to define it.

Let (q, p_a) be a point of Γ_M (so, recall, q is a point of M and p_a is a covariant vector in M at q). Choose a scalar field α on M such that $\nabla_a \alpha$ (a covariant vector field on M) is, at this point q, just p_a. There are, of course, many such scalar fields on M; pick one. Then $\alpha \circ \pi$ is a scalar field on Γ_M. The gradient of this scalar field $\nabla_\beta(\alpha \circ \pi)$ is a covariant vector field on the manifold Γ_M. We evaluate this covariant vector field (on Γ_M) at the point (q, p_a) on Γ_M: this is A_β at the point (q, p_a) of Γ_M. Repeating for each point (q, p_a) of Γ_M, we obtain the desired vector field A_β on Γ_M. This A_β is the thing, in our formulation, which replaces "$p\ dx$" is classical mechanics.

To summarize, an n-dimensional manifold M gives rise, automatically, to a $2n$-dimensional manifold Γ_M, which is mapped smoothly to M, and on which there is a natural covariant vector field A_β.

7. Symplectic Manifolds

As in the previous section, let M be a manifold, and let Γ_M be the cotangent bundle of M. Then, as we have seen, there appears on Γ_M a certain vector field A_β. The curl of this vector field

$$\Omega_{\alpha\beta} = \nabla_\alpha A_\beta - \nabla_\beta A_\alpha \tag{7}$$

is called the **symplectic tensor field** of Γ_M. Which derivative operator should we use in (7)? Fortunately, this is not a decision we have to face, for the right hand side of (7) is independent of the derivative operator used. This is immediate from (6):

$$\nabla'_\alpha A_\beta - \nabla'_\beta A_\alpha = C^\mu{}_{\alpha\beta} A_\mu$$

where, in the last step we have used (5). The right hand side of (7) (except for the factor of two) is called the **exterior derivative** in differential-geometry.

The symplectic field will play a central role in our study of mechanics. The purpose of this section is to discuss its properties.

The most direct way of getting a quick feeling for the symplectic field is to express it in terms of chart. We first do this. Consider a chart on M, writing the coordinates $(x^1, \ldots, x^n)$. Then, as we saw in the previous section, we obtain a chart on Γ_M, with coordinates $(x^1, \ldots, x^n, \kappa_1, \ldots, \kappa_n)$. Now fix a point (q, p_a) of Γ_M, and let the coordinates of this point, with respect to our chart above, be $(\underline{x}^1, \ldots, \underline{x}^n, \underline{\kappa}_1, \ldots, \underline{\kappa}_n)$. We first express A_β, at the point (q, p_a), in terms of this chart on Γ_M. To do this, we just use the definition of A_β. Consider a scalar field α on M which, expressed in terms of our chart on M takes the form

$$\underline{\kappa}_1(x^1 - \underline{x}^1) + \ldots + \underline{\kappa}_n(x^n - \underline{x}^n). \tag{8}$$

It is immediate from the definition of the κ's (pp 18), that $\nabla_a \alpha$, evaluated at q, is just p_a. Furthermore, because of what π does, the scalar field $\alpha \circ \pi$, expressed in terms of our chart on Γ_M, is also given by (8). Hence, $A_\beta|_{(q,p)} = \nabla_\beta(\alpha \circ \pi)|_{(q,p)} = (\underline{\kappa}_1 \nabla_\beta x^1 + \ldots + \underline{\kappa}_n \nabla_\beta x^n)|_{(q,p)}$. Since the point (q, p_a) was arbitrary (except that it had to be in our chart on Γ_M), we have

$$A_\beta = \kappa_1 \nabla_\beta x^1 + \ldots + \kappa_n \nabla_\beta x^n. \tag{9}$$

It is eqn. (9) which suggests we interpret A_β as "$p\ dx$".

We now wish, similarly, to write the symplectic field in terms of these coordinates. This is easy: substitute (9) into (7). Doing this (and using the fact that derivative operators commute when applied to a scalar field), we obtain immediately

$$\Omega_{\alpha\beta} = (\nabla_\alpha \kappa_1)(\nabla_\beta x^1) - (\nabla_\alpha x^1)(\nabla_\beta \kappa_1) + \ldots + (\nabla_\alpha \kappa_n)(\nabla_\beta x^n) - (\nabla_\alpha x^n)(\nabla_\beta \kappa_n) \tag{10}$$

This is the desired formula. In intuitive terms, (10) states that $\Omega_{\alpha\beta}$ has "q-p" components, and "p-q" components, of opposite sign, and that all "q-q" and "p-p" components of $\Omega_{\alpha\beta}$ vanish.

Finally, we introduce three fundamental properties of the symplectic field:

1. $\Omega_{\alpha\beta}$ is antisymmetric, *i.e.*, $\Omega_{\alpha\beta} = -\Omega_{\beta\alpha}$. This is immediate from the definition (7).

2. $\Omega_{\alpha\beta}$ is invertible, *i.e.*, there exists a (unique) antisymmetric tensor field $\Omega^{\alpha\beta}$ such that $\Omega_{\alpha\gamma}\Omega^{\beta\gamma} = \delta_\alpha{}^\beta$, where $\delta_\alpha{}^\beta$, the unit tensor, is defined by the property $\delta_\alpha{}^\beta \xi^\alpha = \xi^\beta$ for all ξ^α. (Invertibility for tensors is the tensor analog of invertibility for matrices.) This property follows, for example, from the explicit expression (10).

3. The curl of $\Omega_{\beta\gamma}$ vanishes, *i.e.*,

$$\nabla_{[\alpha}\Omega_{\beta\gamma]} = 0 \tag{11}$$

Here, and hereafter, square brackets, surrounding tensor indices, mean "write the expression with the enclosed indices in all possible orders, affixing a plus sign to those that are an even permutation of the original order, and a minus sign to those an odd permutation, and, finally, divide by the number of terms". Thus, in detail, the left hand side of (11) means $\frac{1}{6}(\nabla_\alpha\Omega_{\beta\gamma}+\nabla_\beta\Omega_{\gamma\alpha}+\nabla_\gamma\Omega_{\alpha\beta}-\nabla_\alpha\Omega_{\gamma\beta}-\nabla_\beta\Omega_{\alpha\gamma}-\nabla_\gamma\Omega_{\beta\alpha})$. The left hand side of (11) is in fact independent of the choice of derivative operator, an assertion proven exactly, as we proved the same assertion for (7). Finally, that (11) is indeed true can be seen, for example, by substituting (10). (Actually, (11) is a consequence of (7). $\nabla_{[\alpha}\nabla_\beta A_{\gamma]}=0$ for any A_γ. To prove this, note that it is true that when $A_\gamma=\mu\nabla_\gamma V$, and that any A_γ can be expressed as a sum of such terms of this form.)

More generally, a **symplectic manifold** is a manifold (necessarily even dimensional) on which there is specified a tensor field $\Omega_{\alpha\beta}$ having the three properties above.

8. Phase Space

We return to mechanics. Let C be the configuration space of a system. The cotangent bundle of C, Γ_C, is called the **phase space** of the system. It was a consequence of our manner of introducing configuration space that our system, at any instant of time, occupies a particular point of configuration space. It will emerge shortly that our system may also be considered to occupy, at each instant, a point of phase space. That is to say, our system will "possess", at each instant, a pair (q, p_a), where p_a is a vector at point q of C. We shall continue to call q the configuration of the system. The covariant vector p_a at q will be called the **momentum** of the system.

Why does one proceed in this way? At first glance, one might think that the most natural thing to do would be to stay in configuration space. Points of configuration space have "concrete physical significance—they represent actual physical configurations of the actual physical system". Thus, the physics would somehow be closer to the surface in the mathematics if one could carry out the entire description directly in terms of the configuration space. However, a technical difficulty arises. As we pointed out in Sect. 5, knowing only the present configuration of a system does not suffice to determine what the system will do in the future. Roughly speaking, the configuration is half the necessary information. One might, therefore, think of going to the tangent bundle, of describing a system in terms of its configuration q and velocity v^a. Velocity is also, in a sense, "concrete", for one can look at a system (for a short span of time) and determine its velocity. Thus, by going to the tangent bundle, we could work with objects having direct physical significance, and, at the same time, deal with the $2n$ variables (q, v^a) necessary to determine the future behavior of the system.

In short, the tangent bundle seems ideal. Why, then, does one choose the cotangent bundle, the phase space, for the description of our system? The answer is not very satisfactory: because one has, on the cotangent bundle, the symplectic field. This field—this additional structure—seems to be necessary to write the equations which describe mechanics. We shall see that the symplectic field appears in almost every equation. There is nothing on the tangent bundle analogous to the symplectic field. The price one pays for the use of this field is a more tenuous connection between the mathematics and the actual physics. Suppose, for example, that we have an actual physical system, and that we have assigned to it some configuration space C. The system is set into action, and we are permitted to watch the system. We thus obtain a curve in configuration space. What is the momentum of this system at some instant of time? There is no definite way to tell (and, in fact, as we shall see later, there is some ambiguity in this assignment). In a sentence, the momentum of a system is a pure kinematical quantity. On the other hand, the cotangent bundle (phase space) has the same dimension as the tangent bundle. We replace a concrete thing, velocity, by a more abstract thing, momentum, in order to acquire the symplectic field for use in equations. What is unsatisfying is that it is not clear why Nature prefers additional structure over a direct physical interpretation.

Points of phase space will be called **states** of the system. Thus, to give the state of a system, one must specify its configuration and momentum.

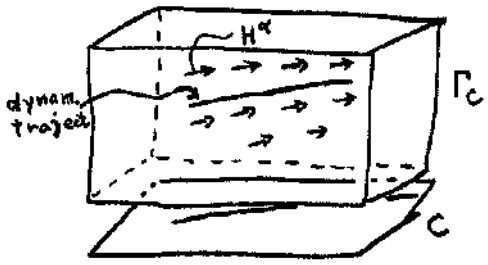
9. The Hamiltonian

The configuration space concerns *kinematics*—what can happen. The phase space is an introduction to the present section, in which we discuss *dynamics*—what actually does happen.

At each instant of time, the system is to possess a state, *i.e.*, a point of phase space, *i.e.*, a configuration and momentum. The evolution of the system is thus described by a curve, $\gamma : \mathbb{R} \to \Gamma_C$, in phase space. Thus, for example, $\pi \circ \gamma$ is a curve in configuration space, a curve which describes the sequence of configurations through which the system passes. Thus, dynamics is to be described by giving a bunch of curves in phase space. Since phase space is $2n$-dimensional, one would expect that precisely one of these curves will pass through each point of phase space. The required statement is this: there is a (smooth) scalar field H on phase space, called the **Hamiltonian**, such that the evolution is described by curves in phase space with tangent vector

$$H^\alpha = \Omega^{\beta\alpha} \nabla_\beta H \tag{12}$$

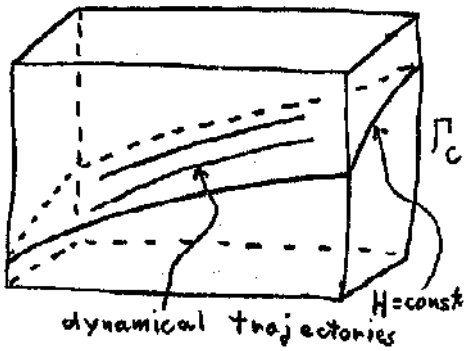
This vector field H^α on phase space is called the **Hamiltonian vector field**. (There is a theorem in differential-geometry to the effect that, given a contravariant vector field on a manifold, there passes, through each point of that manifold, precisely one curve everywhere tangent to that vector field.)



That the statement above has, essentially, the content of **Hamilton's equations** ($\dot{q} = \frac{\partial H}{\partial p}$, $\dot{p} = -\frac{\partial H}{\partial q}$) is clear from the explicit expression (10) for the symplectic tensor field.

The curves in phase space everywhere tangent to the Hamiltonian vector field in phase space will be called the **dynamical trajectories**. The fundamental statement of dynamics can now be formulated as follows: physical systems in Nature seem often to have the property that, when an appropriate configuration space is selected, there exists a Hamiltonian H, a scalar field on phase space, such that the corresponding dynamical trajectories, when projected down to configuration space, give the actual sequence of configurations through which the system passes with time.

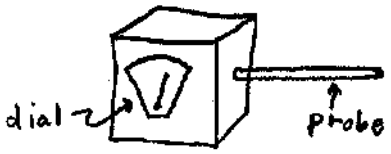
Example: Suppose that our physical system is a free particle (say, to keep things simple, in one dimension) of mass m. One could choose for the Hamiltonian $H = \frac{p^2}{2m}$, and obtain the equations of motion $\dot{x} = \frac{p}{m}$, $\dot{p} = 0$. However, one could as well choose for the Hamiltonian $H = (p + x^2)^2$. Then the equations of motion would be $\dot{x} = 2(p + x^2)$, $\dot{p} = -4x(p + x^2)$. All one actually sees about the physical system is $\ddot{x} = 0$. But this is the case for either Hamiltonian. In particular, if one is given x and $\dot{x}$ for the system at one instant of time, the momentum (and hence the point of phase space) one assigns to the system at that time depends on which Hamiltonian one has chosen. The momentum, in short, is "partially a kinematic and partially a dynamic quantity".



Finally, we show that the scalar field H is constant along the dynamical trajectories. The rate of change of the Hamiltonian along a dynamical trajectory is the directional derivative of H in the direction of the tangent vector to the curve, *i.e.*, the directional derivative of H along the Hamiltonian vector field. But, from (12), $H^\alpha \nabla_\alpha H = (\Omega^{\beta\alpha}\nabla_\beta H)(\nabla_\alpha H) = 0$, where the last step is a consequence of the antisymmetry of $\Omega^{\beta\alpha}$. We can represent this fact geometrically. Draw, in phase space, the $(2n-1)$-dimensional surfaces $H =$ constant. These surfaces fill phase space. The remark above states that each dynamical trajectory remains forever within a single $H =$ constant surface.

10. Observables; Poisson Bracket

By an **observable** of a system, we mean a scalar field on its phase space. Roughly speaking, observables are the things instruments measure. We think of an observable as a box having a dial and a little probe which sticks into the system. The probe is sensitive to what the system is doing, and thus causes the needle to move to some point on the dial. It should be noted, however, that since, given the actual system sitting there, the momentum one assigns to it may depend on the choice (possibly arbitrary to some extent) of a Hamiltonian, one must think of a scalar field on phase space as being associated, not only with the actual instrument which probes the system, but also with our mode of description of the system. Nevertheless, one supposes that the configuration space—phase space—Hamiltonian assignments have been made, somehow, and calls a scalar field on phase space an observable.



Two particular types of observables are of particular interest. Let α be a scalar field on configuration space. Then $\alpha \circ \pi$ is a scalar field on phase space (which ignores momentum, and looks only at configuration). Any observable of this form, $\alpha \circ \pi$, will be called a **configuration observable**. In elementary treatments of particle mechanics, "x", "y", and "z" are configuration observables. Now consider a contravariant vector field ξ^a on configuration space. We introduce the following scalar field on phase space: it assigns, to the point (q, p_a) of phase space, the number obtained by contracting ξ^a at q (a contravariant vector at q) with p_a (a covariant vector at q). We write this scalar field $\xi^a p_a$. An observable expressible in this form will be called a **momentum observable**. In elementary particle mechanics, "p_x", "p_y", and "p_z" are examples of momentum observables (*i.e.*, those defined by unit vectors pointing in the x, y, and z directions). Clearly, some observables (*i.e.*, configuration observables) are "more observable" than others (*e.g.*, momentum observables).

What operations are available on observables? One can certainly add them (addition of scalar fields) and multiply them (outer product of scalar fields). However, there is another, equally important, operation called *Poisson bracket*. If A and B are observables, the **Poisson bracket** of A and B (a third observable) is defined by

$$\{A, B\} = \Omega^{\alpha\beta}(\nabla_\alpha A)(\nabla_\beta B) \tag{13}$$

Note how the symplectic field serves to "mix configuration and momentum components", so that this expression is the same as the familiar one.

There are five basic properties of Poisson brackets:

1. Additive: $\{A + B, C\} = \{A, C\} + \{B, C\}$. This is immediate from (13) and from the fact that the derivatives of (scalar) fields are additive.

2. Leibniz rule: $\{AB, C\} = A\{B, C\} + \{A, C\}B$. This is immediate from the Leibniz rule for derivatives of (scalar) fields.

3. Vanish for constants: If $A =$ constant, then $\{A, B\} = 0$. This is immediate from the fact that the derivative of a constant scalar field is zero.

4. Antisymmetry: $\{A, B\} = -\{B, A\}$. This is immediate from (13) and the fact that $\Omega^{\alpha\beta}$ is antisymmetric.

5. Jacobi identity: $\{A, \{B, C\}\} + \{B, \{C, A\}\} + \{C, \{A, B\}\} = 0$. This is not so immediate. We have

$$\begin{aligned}\{A, \{B, C\}\} &= \Omega^{\alpha\beta}(\nabla_\alpha A)(\nabla_\beta[\Omega^{\gamma\delta}(\nabla_\gamma B)(\nabla_\delta C)\ ]\ ) \\ &= \Omega^{\alpha\beta}[\nabla_\beta\Omega^{\gamma\delta}](\nabla_\alpha A)(\nabla_\gamma B)(\nabla_\delta C) \\ &\quad +\Omega^{\alpha\beta}\Omega^{\gamma\delta}\nabla_\alpha A([\nabla_\beta\nabla_\gamma B]\nabla_\delta C + \nabla_\gamma B[\nabla_\beta\nabla_\delta C]) \quad (14)\end{aligned}$$

where, in the first step, we have substituted (13), and, in the second, we have expanded using the Leibniz rule for a derivative operator. Now suppose we add the right side of (14) to itself three times, switching, each time, the order of A, B, and C as described by the Jacobi identity. Then the second term on the far right in (14) will cancel with itself. Hence, the Jacobi identity will be proven if we can show

$$\Omega^{\beta[\alpha}\nabla_\beta\Omega^{\gamma\delta]} = 0 \quad (15)$$

because this fact would suffice to kill off the first term on the far right in (14). Since $\Omega_{\alpha\beta}$ is invertible, (15) is equivalent to

$$\Omega_{\mu\alpha}\Omega_{\nu\gamma}\Omega_{\sigma\delta}\Omega^{\beta[\alpha}\nabla_\beta\Omega^{\gamma\delta]} = 0 \quad (16)$$

Now differentiate by parts (*i.e.*, use the Leibniz rule for derivative operators):

$$\begin{aligned}&\Omega_{\mu\alpha}\Omega_{\nu\gamma}\Omega_{\sigma\delta}\Omega^{\beta\alpha}\nabla_\beta\Omega^{\gamma\delta} \\ &\quad = \Omega_{\mu\alpha}\Omega_{\nu\gamma}\Omega^{\beta\alpha}\nabla_\beta(\Omega_{\sigma\delta}\Omega^{\gamma\delta}) - \Omega_{\mu\alpha}\Omega_{\nu\gamma}\Omega^{\beta\alpha}\Omega^{\gamma\delta}\nabla_\beta\Omega_{\sigma\delta} \\ &\quad = +\nabla_\mu\Omega_{\sigma\nu}\end{aligned}$$

Finally, using the fact that $\nabla_{[\alpha}\Omega_{\beta\gamma]} = 0$, we obtain (15), and hence the Jacobi identity.

The five properties of the Poisson bracket are easy to remember: the first three are properties of the derivative of scalar fields, the last two are properties of the symplectic field.

Finally, we derive the well-known equation for the rate of change with time of an observable. Let A be an observable. Then the time rate of change of A (along a dynamical trajectory) is (by definition of tangent vector) the directional derivative of A in the direction of the tangent vector to this dynamical trajectory. But this tangent vector is the Hamiltonian vector field. Hence,

$$\dot{A} = H^\alpha\nabla_\alpha A = \Omega^{\beta\alpha}(\nabla_\beta H)(\nabla_\alpha A) = \{H, A\} \quad (17)$$

Note, in particular, that, if we replace "A" by "H" in (17), we obtain, by antisymmetry of the Poisson bracket, $\dot{H} = 0$. We have already seen this.

11. Canonical Transformations

We begin with a little mathematics. Let M and N be manifolds and let $\psi : M \to N$ be a mapping. We say that this ψ is a **diffeomorphism** if

1. ψ is a smooth mapping,
2. ψ^{-1}, a mapping from N to M, exists (*i.e.*, ψ is one-to-one and onto) and is also a smooth mapping.

Intuitively, a diffeomorphism between M and N makes "M and N identical manifolds". Diffeomorphisms are to manifolds as isomorphisms are to vector spaces, or homeomorphisms are to topological spaces.

The remark above suggests that a diffeomorphism from M to N should be prepared to carry tensor fields from M to N and back. This in fact is the case. Let α be a scalar field on M. Then $\alpha \circ \psi^{-1}$ is a scalar field on N. It is sort of the "image" on N of α on M; we write this scalar field $\psi(\alpha)$. We now know how to carry scalar fields from M to N. Since tensor fields are built up in terms of scalar fields, it should now be possible to carry tensor fields from M to N. Indeed, consider the following equations:

$$\psi(\xi^a \nabla_a \alpha) = \psi(\xi^a) \nabla_a \psi(\alpha) \tag{18}$$

$$\psi(\xi^a \eta_a) = \psi(\xi^a) \psi(\eta_a) \tag{19}$$

$$\psi(T^a{}_c{}^{rs}{}_d \xi_a \eta^c \lambda_r \omega_s \kappa^d) = \psi(T^a{}_c{}^{rs}{}_d) \psi(\xi_a) \psi(\eta^c) \psi(\lambda_r) \psi(\omega_s) \psi(\kappa^d) \tag{20}$$

Eqns. (18), (19), and (20) are trying to say that tensor operations commute with the action of ψ. But we wish to use them as the definition of ψ. Eqn. (18) defines $\psi(\xi^a)$, *i.e.*, it defines a mapping from contravariant vector fields on M to contravariant vector fields on N. In words, the definition (18) is this: if ξ^a is a contravariant vector field on M, $\psi(\xi^a)$ is that contravariant vector on N which, applied (via directional derivative) to a scalar field $\psi(\alpha)$ on N, yields the same result as that of applying ξ^a to α (on M), and carrying the result over (via ψ) to N. In short, (18) requires that the operation "directional derivative" be invariant under the pushing of fields from M to N. Similarly, (19) defines $\psi(\eta_a)$, *i.e.*, (19) defines a mapping from covariant vector fields on M to covariant vector fields on N. This definition amounts to requiring that contraction be invariant under the "pushing action" of ψ. Finally, (20) defines a mapping from arbitrary tensor fields on M to fields on N.

In short, a diffeomorphism from one manifold to another provides a mechanism for carrying tensor fields from the first manifold to the second. This mechanism is completely and uniquely determined by the following properties: on scalar fields, it is the obvious thing, and it commutes with all tensor operations.

Now consider a system, with configuration space C, phase space Γ_C, symplectic structure $\Omega_{\alpha\beta}$, and Hamiltonian H. A diffeomorphism $\psi : \Gamma_C \to \Gamma_C$ from phase space to itself which leaves invariant the symplectic field,

$$\psi(\Omega_{\alpha\beta}) = \Omega_{\alpha\beta} \tag{21}$$

is called a **canonical transformation**. That is to say, canonical transformations are mappings from phase space to itself which preserve all of the structure of interest there (smoothness, symplectic). Note that canonical transformations need not leave the Hamiltonian invariant. When this is the case, *i.e.*, when

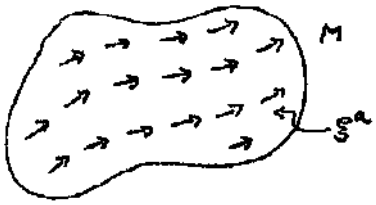
$$\psi(H) = H \tag{22}$$

the canonical transformation will be called a **symmetry**. For a moment, regard the phase space, and symplectic field as representing "kinematics", and

these things, together with the Hamiltonian, as representing "dynamics". Then the following sentence is both a motivation for, and a summary of, the definitions above: canonical transformations are to kinematics as symmetries are to dynamics as isomorphisms are to vector spaces.

Because they are difficult to manipulate, one seldom uses canonical transformations in practice. Instead, one introduces what are called *infinitesimal canonical transformations*. In order to define infinitesimal canonical transformations, we must, again, introduce a few properties of manifolds.

Let M be a manifold. Consider the identity diffeomorphism on M, the one that takes each point of M to itself. Then, roughly speaking, an infinitesimal diffeomorphism on M should be the one which is near the identity, *i.e.*, one that takes each point of M to a nearby point. But a contravariant vector field on M defines a "direction of motion in M" at each point. These remarks suggest that a contravariant vector field on M is the precise geometrical object appropriate to replace the intuitive notion of an "infinitesimal diffeomorphism". A diffeomorphism from M to M carries tensor fields on M to tensor fields on M. One would expect, therefore, that an infinitesimal diffeomorphism on M carries tensor fields on M to nearby tensor fields on M. The appropriate thing to consider, therefore, is the "infinitesimal change in the tensor fields under infinitesimal diffeomorphism". The precise representation of an "infinitesimal diffeomorphism" on M is a contravariant vector field on M. This remark suggests that the precise representation of "the infinitesimal change in a tensor field under an infinitesimal diffeomorphism on M" should be something like "the directional derivative of a tensor field in the direction of ξ^a". The remarks above are merely motivation for the definitions which follow.



Fix a contravariant vector field ξ^a on the manifold M. The symbol $\pounds_\xi$ will be called the **Lie derivative** (in the direction of ξ^a). For α, a scalar field on M, set

$$\pounds_\xi \alpha = \xi^a \nabla_a \alpha. \tag{23}$$

Thus, the Lie derivative of a scalar field is the directional derivative. As before, we now wish to extend the notion of the Lie derivative from scalar fields to tensor fields. Consider the following equations:

$$\pounds_\xi(\eta^a \nabla_a \alpha) = (\pounds_\xi \eta^a)\nabla_a \alpha + \eta^a \nabla_a(\pounds_\xi \alpha) \tag{24}$$

$$\pounds_\xi(\eta^a \lambda_a) = (\pounds_\xi \eta^a)\lambda_a + \eta^a \pounds_\xi \lambda_a \tag{25}$$

$$\begin{aligned} \pounds_\xi(T^{ac}{}_d \eta_a \lambda_c \kappa^d) \quad &= \quad \pounds_\xi(T^{ac}{}_d)\eta_a \lambda_c \kappa^d + T^{ac}{}_d \pounds_\xi(\eta_a)\lambda_c \kappa^d \\ &\quad + T^{ac}{}_d \eta_a \pounds_\xi(\lambda_c)\kappa^d + T^{ac}{}_d \eta_a \lambda_c \pounds_\xi(\kappa^d) \end{aligned} \tag{26}$$

These equations are trying to say that the Lie derivative satisfies the "Leibniz rule, as any good directional derivative type operator should do". But we use

them as definitions. Eqn. (24) is the definition of the Lie derivative of a contravariant vector field (η^a) on M; Eqn. (25) the definition of the Lie derivative of a covariant vector field on M; Eqn. (26) the definition of the Lie derivative of an arbitrary tensor field on M. To summarize, the Lie derivative, "the directional derivative generalized to tensor fields", is completely and uniquely characterized by the following properties: on scalar fields, it is the obvious thing, and it satisfies Leibniz-type rules.

One actually uses in practice, not the definitions above for the Lie derivative, but more explicit expressions we now derive. We have

$$\xi^b\nabla_b(\eta^a\nabla_a\alpha) = \pounds_\xi\eta^a(\nabla_a\alpha) + \eta^a\nabla_a(\xi^b\nabla_b\alpha)$$

$$(\xi^b\nabla_b\eta^a)\nabla_a\alpha + \xi^b\eta^a\nabla_b\nabla_a\alpha = (\pounds_\xi\eta^a)\nabla_a\alpha + (\eta^a\nabla_a\xi^b)\nabla_b\alpha + \eta^a\xi^b\nabla_a\nabla_b\alpha$$

$$(\xi^b\nabla_b\eta^a)\nabla_a\alpha = (\pounds_\xi\eta^a)\nabla_a\alpha + (\eta^a\nabla_a\xi^b)\nabla_b\alpha$$

where the first equation results from using (23) in (21), the second equation results from expanding using the Leibniz rule (for a derivative operator), and the third equation results from the fact that derivatives commute on scalar fields. Thus, since α is arbitrary in the last equation above, we have

$$\pounds_\xi\eta^a = \xi^b\nabla_b\eta^a - \eta^b\nabla_b\xi^a \tag{27}$$

We now, similarly, use (23) and (27) in (25)

$$\xi^b\nabla_b(\eta^a\lambda_a) = (\xi^b\nabla_b\eta^a - \eta^b\nabla_b\xi^a)\lambda_a + \eta^a\pounds_\xi\lambda_a$$

$$(\xi^b\nabla_b\eta^a)\nabla_c\alpha + \xi^b\eta^a\nabla_b\nabla_a\alpha = (\pounds_\xi\eta^a)\nabla_a\alpha + (\eta^a\nabla_a\xi^b)\nabla_b\alpha + \eta^a\xi^b\nabla_a\nabla_b\alpha$$

$$(\xi^b\nabla_b\eta^a)\nabla_a\alpha = (\pounds_\xi\eta^a)\nabla_a\alpha + (\eta^a\nabla_a\xi^b)\nabla_b\alpha$$

Since η^a is arbitrary,

$$\pounds_\xi\lambda_a = \xi^b\nabla_b\lambda_a + \lambda_b\nabla_a\xi^b \tag{28}$$

Finally, using (27) and (28) in (26), we obtain, in precisely the same way above,

$$\begin{aligned}\pounds_\xi T^{ab}{}_{cd} &= \xi^m\nabla_m T^{ab}{}_{cd} - T^{mb}{}_{cd}\nabla_m\xi^a - T^{am}{}_{cd}\nabla_m\xi^b \\ &\quad + T^{ab}{}_{md}\nabla_c\xi^m + T^{ab}{}_{cm}\nabla_d\xi^m\end{aligned} \tag{29}$$

We could just as well have defined the Lie derivative operation by (29). (Note that (23), (27), and (28) are just special cases, and that (24), (25), and (26) follow immediately from (29).) Eqn. (29) is easily the most useful expression for the Lie derivative. Finally, we note that the right hand side of (29) is in fact independent of the choice of derivative operator. This is obvious from our definition, but can also be checked directly from (6).

We now again return to our system, with its configuration space, phase space, symplectic structure, and Hamiltonian. An **infinitesimal canonical transformation** on the system is a contravariant vector field ξ^a on phase space, Γ_C, such that

$$\pounds_\xi\Omega_{\alpha\beta} = 0 \tag{30}$$

An **infinitesimal symmetry** is an infinitesimal canonical transformation ξ^a which, in addition, satisfies

$$\pounds_\xi H = 0 \tag{31}$$

Everything in this section so far, except for (29) and these two definitions, can be regarded as motivation.

There is a crucial formula, which simplifies the discussion of infinitesimal canonical transformations, namely,

$$3\xi^\gamma\nabla_{[\alpha}\Omega_{\beta\gamma]} - \pounds_\xi\Omega_{\alpha\beta} - 2\nabla_{[\alpha}(\Omega_{\beta]\gamma}\xi^\gamma) = 0 \tag{32}$$

To prove (32), we simply expand it

$$\begin{aligned} &\xi^\gamma \nabla_\alpha \Omega_{\beta\gamma} + \xi^\gamma \nabla_\beta \Omega_{\gamma\alpha} + \xi^\gamma \nabla_\gamma \Omega_{\alpha\beta} \\ -&\xi^\gamma \nabla_\gamma \Omega_{\alpha\beta} - \Omega_{\gamma\beta} \nabla_\alpha \xi^\gamma - \Omega_{\alpha\gamma} \nabla_\beta \xi^\gamma \\ &-\nabla_\alpha(\Omega_{\beta\gamma}\xi^\gamma) + \nabla_\beta(\Omega_{\alpha\gamma}\xi^\gamma) \quad = \quad 0 \end{aligned}$$

The first term of (32) gives the first three terms above, the second term of (32) the next three terms, and the last term of (32) the last two terms. But the expansion above holds (the third term cancels the fourth, the first, fifth, and seventh cancel, and the second, sixth, and eighth cancel). Hence (32) is true.

Since the curl of the symplectic field vanishes (Eqn. 11), (32) becomes

$$\pounds_\xi \Omega_{\alpha\beta} + 2\nabla_{[\alpha}(\Omega_{\beta]\gamma}\xi^\gamma) = 0 \tag{33}$$

Thus, ξ^α is an infinitesimal canonical transformation if and only if the curl of $\Omega_{\beta\gamma}\xi^\gamma$ vanishes. But (at least, locally) the curl of a covariant vector field on a manifold vanishes if and only if it is a gradient. Thus, ξ^α is an infinitesimal canonical transformation if and only if $\Omega_{\beta\gamma}\xi^\gamma = \nabla_\beta A$ for some scalar field A on phase space. Contracting this equation with $\Omega_{\beta\alpha}$, we find: ξ^α is an infinitesimal canonical transformation if and only if $\xi^\alpha = \Omega^{\beta\alpha}\nabla_\beta A$ for some A. This A is called the **generator** (or **generating function**) of the infinitesimal canonical transformation.

Thus, observables generate infinitesimal canonical transformations. It should also be clear why one often says that "the evolution of the system is the unfolding of a sequence of infinitesimal canonical transformations". (We would say: "The Hamiltonian vector field is an infinitesimal canonical transformation.")

Thus, infinitesimal canonical transformations can be described in terms of scalar fields on phase space rather than vector fields (an enormous simplification). What condition must an observable A satisfy in order that the infinitesimal canonical transformation it generates be an infinitesimal symmetry? This is easy to answer. We ask that the Lie derivative of H in the direction of $\Omega^{\beta\alpha}\nabla_\beta A$ vanish. That is to say, we ask that $(\Omega^{\beta\alpha}\nabla_\beta A)(\nabla_\alpha H) = 0$. That is to say, we ask that the Poisson bracket, $\{A, H\} = 0$. But, from (17), this is precisely the statement that A be constant along each dynamical trajectory, *i.e.*, that A be a constant of the motion. To summarize, we have shown that every constant of the motion generates an infinitesimal symmetry, and that, conversely, every symmetry is generated by a constant of the motion.

12. Algebraic Observables

Recall that the first step in the description of the system was the assignment to that system of a configuration space. This configuration space then led to phase space with its symplectic structure. Note, however, that essentially all of the notions of mechanics relate directly to phase space rather than configuration space. That is to say, the primary purpose in introducing configuration space has been to obtain phase space. In fact, one could just as well have treated mechanics as follows:

A system is described by a certain symplectic manifold, called phase space, on which there is a certain scalar field, called the Hamiltonian. One then has dynamical trajectories, Poisson brackets, canonical transformations, etc.

The present section is devoted to certain constructions which use, in an essential way, the presence of configuration space.

In sect. 10, we introduced configuration and momentum observables. We wish to generalize this notion. Consider a tensor field $R^{a_1\cdots a_r}$, with r indices, on configuration space, and suppose that this tensor field is symmetric. (That is, suppose $R^{a_1\cdots a_r} = R^{(a_1\cdots a_r)}$, where round brackets, surrounding a collection of tensor indices, mean "add the expression to all expressions obtained by changing the order of the surrounded indices, and divide by the number of terms".) We associate, with this tensor field, an observable. Let (q, p_a) be a point of phase space. Then, $R^{a_1\cdots a_r}(q)p_{a_1}\cdots p_{a_r}$ is a number. Thus, we have assigned a number to each point of phase space, *i.e.*, we have defined an observable. We shall write this observable $R^{a_1\cdots a_r}p_{a_1}\cdots p_{a_r}$ or sometimes as $O(R)$. For $r = 0$ (*i.e.*, when $R^{a_1\cdots a_r}$ has no indices), we obtain a configuration variable; for $r = 1$, a momentum observable. Note, however, that we also obtain observables which are quadratic, cubic, *etc.*, in the momenta.

Our goal is to express, in terms of the tensor fields, the results on various operations on these observables.

Evidently, one can add these totally symmetric tensor fields (*i.e.*, $R^{a_1\cdots a_r}$ and $S^{a_1\cdots a_s}$) when they have the same number of indices (*i.e.*, when $r = s$). There is also a product. If $R^{a_1\cdots a_r}$ and $S^{a_1\cdots a_s}$ are two such fields, we define their product, written $R\cap S$ (indices suppressed), by $R^{(a_1\cdots a_r}S^{a_{r+1}\cdots a_{r+s})}$. Note that the number of indices in the product is $r + s$. Obviously, the product is also associative and commutative, and distributive over addition. Finally, we define a bracket relation on such tensor fields. Let $[R, S]$ (indices suppressed) be the tensor field

$$rR^{ma_1\cdots a_{r-1}}\nabla_m S^{a_r\cdots a_{r+s-1}} - sS^{ma_1\cdots a_{s-1}}\nabla_m R^{a_s\cdots a_{r+s-1}} \tag{34}$$

First note, from (6), that this expression is independent of the choice of derivative operator. Next, note that the bracket is antisymmetric, additive, and satisfies the Leibniz rule: $[R, S] = -[S, R]$, $[R + S, T] = [R, T] + [S, T]$, $[R\cap S, T] = R\cap[S, T] + [R, T]\cap S$. It is also true (and it can be verified directly from (34)) that this bracket relation satisfies the Jacobi identity: $[[R, S], T] + [[S, T], R] + [[T, R], S] = 0$. Finally, note that, when $r = 1$, (*i.e.*, when R is a vector), our bracket is just the Lie derivative $[R, S] = \pounds_R S$.

We now claim: the three operations (sum, product, bracket) introduced above on symmetric tensor fields on configuration space correspond precisely to the three operations (sum, product, Poisson bracket) on the corresponding observables. That is to say,

$$O(R + S) = O(R) + O(S) \tag{35}$$

$$O(R\cap S) = O(R)O(S) \tag{36}$$

$$O([R, S]) = \{O(R), O(S)\} \tag{37}$$

We first note that (35) and (36) are obvious. To prove (37), one first checks, from (10), that it is valid when only configuration and momentum observables are permitted. But any tensor field can be written as a sum of outer products of vectors. Inserting such sums in (37), using the Leibniz rule (on the left, for the bracket; on the right, for the Poisson bracket) and the fact that (37) holds for vectors, we obtain (37) in general. This proves our assertion.

Thus, all the operations on observables yields algebraic observables (*i.e.*, those obtained from tensor fields on configuration space) when applied to algebraic observables. In fact, nearly every observable one normally deals with in mechanics is algebraic. An example follows.

Consider the Hamiltonian $H = g^{ab}p_a p_b + A^a p_a + V$, where g^{ab}, A^a, and V are fields on configuration space (with, of course, g^{ab} symmetric). The first term, for example, could be the Hamiltonian of a free particle in Euclidean space, where g^{ab} is the usual metric of Euclidean space. The first and third terms could be the Hamiltonian of a particle in a potential (V). The first and second terms could be the Hamiltonian of a charged particle (where A^a is a vector potential for the magnetic field). There appear, in fact, to be very few, if any, systems whose Hamiltonian cannot be put into this form. Let us ask when a momentum observable, $\xi^a p_a$, is a constant of motion for this Hamiltonian. The condition is that the Poisson bracket of the Hamiltonian and the observable vanish. Thus, we require that $\{\xi, g\} = 0$, $\{\xi, A\} = 0$, $\{\xi, V\} = 0$. That is to say, we require

$$\pounds_\xi g^{ab} = 0, \qquad \pounds_\xi A^b = 0, \qquad \pounds_\xi V = 0.$$

But this is just the statement that the "infinitesimal diffeomorphism" on configuration space generated by ξ^a leave invariant the various tensor fields (on configuration space) which go into the Hamiltonian.

Suppose our system is a free particle, so $A^a = 0$ and $V = 0$. Then, $\xi^a p_a$ is a constant of the motion whenever $\pounds_\xi g^{ab} = 0$, where g^{ab} is the metric of Euclidean space. There are six such vector fields on Euclidean space, representing translations and rotations. The constants of motion associated with translations are what are usually called *momenta*. The constants of motion associated with the rotations are what are usually called *angular momenta*. Now suppose we introduce a potential V, and ask which of our constants of motion remain constants of the motion. It is precisely those which also satisfy $\pounds_\xi V = \xi^a \nabla_a V = 0$. In other words, it is the symmetries of Euclidean space which also leave invariant the potential. If, for example, V is spherically symmetric, there will be three such vector fields, ξ_1^a, ξ_2^a, and ξ_3^a (namely, those which represent rotations about the origin of spherical symmetry of V). Set $J = \xi_1^a \cap \xi_1^a + \xi_2^a \cap \xi_2^a + \xi_3^a \cap \xi_3^a$, a symmetric, second-rank tensor field on configuration space. By the Jacobi identity, $O(J)$ is also a constant of the motion (a quadratic one). This constant of the motion is called the *squared angular momentum* (for a spherically symmetric system). The introduction of an A^a destroys, in general, these constants of the motion—unless, of course, the A^a happens to also be invariant under the corresponding infinitesimal diffeomorphism on configuration space.

Finally, note that constants of the motion remain constants of the motion under addition, product, and Poisson brackets. This is true, in particular, of algebraic constants of the motion.

Part III.
Quantum Mechanics

13. Introduction

Mechanics is pretty, tidy, and natural looking. It suffers, however, from one serious flaw: Nature does not seem to behave that way. Specifically, quantum mechanics has been found to represent a more accurate description of Nature than classical mechanics. Roughly speaking, quantum mechanics is a "smeared out version of classical mechanics, where the amount of smearing is governed by a certain constant, Planck's constant $\hbar$." This last statement is, even as a "roughly speaking", somewhat unsatisfactory. It suggests that the principles of quantum mechanics are somehow to be imposed on a classically defined system, that quantum mechanics is some sort of thin veneer over classical mechanics. Thus, one often speaks of "quantizing" a classical system. (The situation is perhaps analogous to that in which one first treats matter as a continuum, and then "atomizes" it.) Systems, apparently, "really are" quantum systems—classical mechanics is merely a simplified approximation for the limit in which one is insensitive to the "smearing". Nonetheless, it is common, presumably because classical mechanics makes good contact with everyday life, to regard quantum mechanics as some sort of correction to classical mechanics.

Our goal is to quantize the system discussed in part II, *i.e.*, to write down the quantum description of various systems, and check that this description reduces, in an appropriate limit, to the classical description. It would perhaps be more logical, as the above remark suggests, to do part III before part II. We have artificially reversed the order to provide motivation: part II can be motivated from everyday life, and part III can be motivated from part II.

There are at least three essentially equivalent formulations of quantum mechanics: *Schrodinger, Heisenberg, and algebraic*. Essentially equivalent formulations, however, need not necessarily be of equal value: one may give better insight, be more readily adapted to other contexts, or suggest exotic generalizations, than another. The view is not uncommon that the Schrodinger formulation of quantum mechanics is less natural than the others. The reason is that the Schrodinger formulation seems to fit badly with the principle of relativity. Nonetheless, the Schrodinger formulation does tie in nicely with classical mechanics, and so we shall begin with this point of view.

Let us, for the moment, regard classical mechanics as consisting of states, observables, and dynamics. The states, of course, are the points of phase space, the observables scalar fields on phase space. We can regard the states and observables as the kinematics. The dynamics consists of the Hamiltonian dynamical trajectories, *etc.*

Each of these three classical notions has an analog in quantum mechanics. The space of states in classical mechanics is to be replaced by a corresponding (although different) space of states in quantum mechanics. In either case, the system is to be regarded as possessing, at each instant of time, one of its possible states. Dynamics plays essentially the same role in quantum mechanics as in classical mechanics: it is the means by which one describes the successive states through which a system passes with time. It is with the observables, however, that a significant departure between classical and quantum mechan-

ics, apparently, takes place. Classically, an observable assigns a number to each state—a feature one might expect to be characteristic of what one means by a "measuring instrument". It might seem reasonable, therefore, to suppose that observables also in quantum mechanics will have this same characteristic property. However, observables emerge initially in quantum mechanics as mappings from states to states. This mathematical feature is presumably a reflection of the following physical fact: *observations on a classical system can be performed with negligible effect on the system observed, while this is not the case for a quantum system.* In other words, the mere assignment of a number to each state is not enough information to define a quantum observable; one must also give information regarding the state into which the system will be thrown by this act of observation. This "extra information" is carried by having observables map states to states rather than to real numbers.

14. Densities. Integrals.

The things which can be integrated over a manifold, it turns out, are, not scalar fields, but things called *scalar densities*. Scalar densities, in turn, represent a special case of *tensor densities*. We introduce here the appropriate definitions and properties.

Let M be an n-dimensional manifold. An **alternating tensor**, $\epsilon^{a_1\cdots a_n}$, at a point of M is an n-index contravariant tensor at that point which is antisymmetric, *i.e.*, which reverses sign under interchange of any two of its indices, *i.e.*, which satisfies $\epsilon^{[a_1\cdots a_n]} = \epsilon^{a_1\cdots a_n}$. If ϵ and ϵ' are two alternating tensors at the same point, with the second nonzero, then there exists a number α such that $\epsilon^{a_1\cdots a_n} = \alpha\epsilon'^{a_1\cdots a_n}$. (Proof: Choose a basis $v_{1a}, \ldots, v_{na}$, for the covariant vectors at the point, and note that, by antisymmetry, an alternating tensor is completely and uniquely determined by the value of the number $\epsilon^{a_1\cdots a_n} v_{1a_1}\cdots v_{na_n}$.) Thus, the alternating tensors at a point of a manifold form a one-dimensional vector space. An **alternating tensor field** is an n-index, contravariant, antisymmetric tensor field on M, *i.e.*, an assignment (smoothly) of an alternating tensor to each point of M.

Fix a real number s. A **tensor density**, *e.g.*, $T^p{}_{bc}{}^w$, at a point of M is a mapping from alternating tensors $\epsilon^{a_1\cdots a_n}$ at that point to tensors $T^p{}_{bc}{}^w(\epsilon)$ at that point such that

$$T^p{}_{bc}{}^w(\alpha\epsilon) = (\alpha)^s T^p{}_{bc}{}^w(\epsilon) \tag{38}$$

for any real number α. The number s is called the **weight** of the tensor density. Since any two alternating tensors are proportional, (38) implies that a knowledge of $T^p{}_{bc}{}^w(\epsilon)$ for one nonzero alternating tensor $\epsilon^{a_1\cdots a_n}$ determines $T^p{}_{bc}{}^w(\epsilon)$ for all ϵ, and hence, determines the tensor density $T^p{}_{bc}{}^w$. Thus, a choice of alternating tensor at a point defines a one-to-one correspondence between tensor densities at that point and tensors at that point. Roughly speaking, a tensor density is a tensor having certain "scaling behavior" with alternating tensors. Note, from (38), that a tensor density of weight zero is independent of the alternating tensor. That is to say, tensor densities of weight zero are just ordinary tensors.

Example: Let $\kappa_{a_1\cdots a_n}$ be a tensor at a point of M. For any alternating tensor $\epsilon^{a_1\cdots a_n}$ at this point, set $\phi(\epsilon) = \kappa_{a_1\cdots a_n}\epsilon^{a_1\cdots a_n}$. Thus, we have defined a mapping from alternating tensors at a point to scalars at that point (*i.e.*, to real numbers). It is immediate that this mapping satisfies (38) with $s = 1$. Hence, $\phi(\epsilon)$ is a scalar density of weight $+1$. More generally, since tensor densities of weight $+1$ define, via (38), linear mappings from alternating tensors to tensors, every tensor density of weight $+1$ can be represented, as above, by a tensor. Similarly, $\lambda_{a_1\cdots a_n b_1\cdots b_n}\epsilon^{a_1\cdots a_n}\epsilon^{b_1\cdots b_n}$ is a scalar density of weight $+2$. It is now clear that every tensor density of positive integral weight can be represented as a tensor. Now let $\eta^{a_1\cdots a_n}$ be a nonzero alternating tensor at a point of M. Then, for $\epsilon^{a_1\cdots a_n}$ another alternating tensor at that point, set $\chi(\epsilon) = \alpha$, where α is the number such that $\eta^{a_1\cdots a_n} = \alpha\epsilon^{a_1\cdots a_n}$. Clearly, this $\chi(\epsilon)$ (a mapping from alternating tensors to scalars) is a scalar density of weight -1. Thus, a nonzero alternating tensor defines a scalar density of weight -1. Similarly, every tensor density of negative integral weight can be represented by a tensor. The purpose of introducing tensor densities is to enable us to deal with non-integral weights.

A **tensor density field** on M is a (smooth) assignment of a tensor density to each point of M, where the weight is the same for each point of M. (Here, of course, "smooth" means that a smooth tensor field results from a smooth alternating tensor field.)

Which of our tensor operations are applicable to tensor density fields? Let $T^a{}_{cd}$ and $T'^a{}_{cd}$ be tensor density fields, and, for any alternating tensor ϵ, set $(T^a{}_{cd} + T'^a{}_{cd})(\epsilon) = T^a{}_{cd}(\epsilon) + T'^a{}_{cd}(\epsilon)$. Now let T and T' have weights s and s', respectively. Then, clearly, the right hand side of this equation is a tensor density (*i.e.*, the right hand side satisfies (38)) only when $s = s'$. Thus, the sum of two tensor density fields is defined precisely when they have the same

weight (and, of course, the same index structure). The weight of the sum is the weight of each addend. Similarly, (*e.g.*, $(T^a{}_{cd}W_{pq})(\epsilon) = T^a{}_{cd}(\epsilon)W_{pq}(\epsilon)$), the outer product of any two tensor density fields is defined. The weight of the product is the sum of the weights of the factors. (Note that, for weight zero, *i.e.*, for ordinary tensor fields, these operations reduce to ordinary sum and outer product.) Finally, index substitution and contraction are defined in the obvious way for tensor densities, and, when applied, yield a density of the same weight. In short, except for the provision that addition is defined only between densities having the same weight, all operations on tensors extend to operations on tensor densities.

Fix a derivative operator ∇_a on M, so this ∇_a acts on tensor fields. We wish to extend this action to tensor densities. Let $\epsilon^{a_1\cdots a_n}$ be an alternating tensor field on M. Then $\nabla_a\epsilon^{a_1\cdots a_n} = \lambda_a\epsilon^{a_1\cdots a_n}$ for some covariant vector field λ_a. (Proof: the left hand side of this equation is antisymmetric in $a_1, \cdots, a_n$. But two alternating tensors are proportional.) We now define the derivative of a tensor density, *e.g.*, $T^b{}_{cd}$, by

$$(\nabla_a T^b{}_{cd})(\epsilon) = \nabla_a(T^b{}_{cd}(\epsilon)) - s\lambda_a T^b{}_{cd}(\epsilon) \tag{39}$$

where s is the weight of $T^b{}_{cd}$. That is to say, $\nabla_a T^b{}_{cd}$ is to be the tensor density field whose action on any alternating tensor field ϵ is given by the right hand side of (39). We must check that the right hand side of (39) indeed satisfies (38). To this end, let $\epsilon'^{a_1\cdots a_n} = \alpha\epsilon^{a_1\cdots a_n}$. Then $\nabla_a(\epsilon'^{a_1\cdots a_n}) = \nabla_a(\alpha\epsilon^{a_1\cdots a_n}) = \alpha\nabla_a(\epsilon^{a_1\cdots a_n}) + (\nabla_a\alpha)\epsilon^{a_1\cdots a_n}$. Hence, $\lambda'_a = \lambda_a + \alpha^{-1}(\nabla_a\alpha)$. We then have, for the right side of (39),

$$\begin{aligned}
\nabla_a(T^b{}_{cd}(\epsilon')) - s\lambda'_a T^b{}_{cd}(\epsilon') &= \nabla_a(T^b{}_{cd}(\alpha\epsilon)) - s(\lambda_a + \alpha^{-1}(\nabla_a\alpha))T^b{}_{cd}(\alpha\epsilon) \\
&= \nabla_a(\alpha^s T^b{}_{cd}(\epsilon)) - s(\lambda_a + \alpha^{-1}(\nabla_a\alpha))\alpha^s T^b{}_{cd}(\epsilon) \\
&= [\alpha^s\nabla_a(T^b{}_{cd}(\epsilon)) + (s\alpha^{s-1}(\nabla_a\alpha))T^b{}_{cd}(\epsilon)] \\
&\qquad -s\lambda_a\alpha^s T^b{}_{cd}(\epsilon) - s\alpha^{s-1}(\nabla_a\alpha)T^b{}_{cd}(\epsilon) \\
&= \alpha^s\nabla_a(T^b{}_{cd}(\epsilon)) - s\lambda_a(\alpha^s)T^b{}_{cd}(\epsilon) \\
&= \alpha^s[\nabla_a(T^b{}_{cd}(\epsilon)) - s\lambda_a T^b{}_{cd}(\epsilon)]
\end{aligned}$$

Thus, (39) indeed defines a tensor density field of weight s. To summarize, a derivative operator on tensor fields extends, via (39), to an operator on tensor density fields, where taking the derivative does not change the weight. Note, from (39), that a derivative operator on tensor density fields is additive and satisfies the Leibniz rule.

My n^{th}-generation copy of this page had unreadable sections. So, this page is my attempt to reconstruct the content. I hope that I have maintained the style of the presentation and that I did not introduce any errors. (Please check!)

Now suppose that we have two derivative operators, ∇_a and ∇'_a, on M, related via (6). We wish to find the expression, analogous to (6), relating the actions of these operators on tensor densities. Let $\epsilon^{a_1\cdots a_n}$ be an alternating tensor field on M. Then, from (6), $\nabla'_a(\epsilon^{a_1\cdots a_n}) = \nabla_a(\epsilon^{a_1\cdots a_n}) - C^m{}_{am}\epsilon^{a_1\cdots a_n}$. That is to say, $\lambda'_a = \lambda_a - C^m{}_{am}$. Then, (39) implies $(\nabla'_a T^b{}_{cd} - \nabla_a T^b{}_{cd})(\epsilon) = (\nabla'_a - \nabla_a)T^b{}_{cd}(\epsilon) + sC^m{}_{am}T^b{}_{cd}(\epsilon)$. Since, using (6), we know how $(\nabla'_a - \nabla_a)$ acts on a tensor $T^b{}_{cd}(\epsilon)$, we can expand the first term on the right hand side to obtain $(\nabla'_a T^b{}_{cd} - \nabla_a T^b{}_{cd})(\epsilon) = -C^b{}_{am}T^m{}_{cd}(\epsilon) + C^m{}_{ac}T^b{}_{md}(\epsilon) + C^m{}_{ad}T^b{}_{cm}(\epsilon) + sC^m{}_{am}T^b{}_{cd}(\epsilon)$. Thus, the generalization of (6) is

$$\begin{aligned}
(\nabla'_a T^b{}_{cd} - \nabla_a T^b{}_{cd}) &= -C^b{}_{am}T^m{}_{cd} + C^m{}_{ac}T^b{}_{md} + C^m{}_{ad}T^b{}_{cm} \\
&\qquad + sC^m{}_{am}T^b{}_{cd}
\end{aligned} \tag{40}$$

In short, things about tensor fields carry over with appropriate (and minor) modifications to things about tensor density fields.

Our final task is to introduce integrals. At a point of the n-dimensional manifold M, choose N contravariant vectors, $\xi^a_1, \ldots, \xi^a_n$. Since a contravariant vector can be regarded as defining an"infinitesimal motion at a point", these n vectors can be regarded as defining an infinitesimal parallelepiped. We would like to associate with this figure a "volume". If we were to multiply any of the vectors $\xi^a_1, \ldots, \xi^a_n$ by a factor, the volume should be multiplied by that factor. If we add to any of these vectors a multiple of another vector, the volume should remain the same. Thus, the volume of this infinitesimal figure should be proportional to the alternating tensor $\xi^{[a_1}_1 \cdots \xi^{a_n]}_n$. We can think of a volume element as assigning to each infinitesimal parallelepiped, *i.e.*, to each alternating tensor, a number, linearly in the alternating tensor. In other words, a volume element should be a scalar density of weight $+1$. It is a scalar density of weight $+1$ which can be integrated over a manifold to obtain a number.

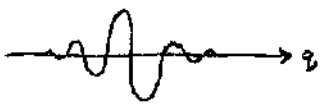
To actually define an integral over a manifold, we must go back to charts. Let ϕ be a scalar density of weight $+1$. Choose a chart, with coordinates $x^1, \ldots, x^n$. Consider the unique alternating tensor satisfying $\epsilon^{a_1 \cdots a_n}(\nabla_{a_1} x^1) \cdots (\nabla_{a_n} x^n) = 1$. Then the integral of ϕ over the region covered by this chart is, by definition, the integral $\int \phi(\epsilon) dx^1 \cdots dx^1$. It is because our original density was of weight $+1$ that this integral is independent of the choice of chart.

15. States

In classical mechanics, the states are points of phase space. In quantum mechanics, the **states** are complex-valued densities ψ on configuration space C, of weight $+\frac{1}{2}$, for which the integral

$$\int_C \bar{\psi}\psi = \|\psi\|^2 \tag{41}$$

converges (*i.e.*, is finite), where a bar denotes the complex conjugate. There is a sense in which a quantum state can be regarded as a "smeared out" classical state. The "configuration half", q, of a classical state (q, p_a), is represented by $\bar{\psi}\psi$, a density of weight $+1$. The density could, for example, be "peaked" near a point q of configuration space. The "lack of peaking" corresponds to the "amount of smearing out". The momentum information in a classical state is represented, within a quantum state, by the phase of the density ψ. The "direction in which the phase is changing fastest" replaces the "direction of p_a", while the rate of change of the phase represents the "magnitude of p_a". Thus, the two "point pieces of information", q and p_a, which go into a classical state are reflected, in a smeared out way, by the real and imaginary parts of the density ψ.



There is one remarkable thing that has happened in the transition from classical to quantum states. Recall that, in classical mechanics, the introduction of configuration space was essentially for motivational reasons. One could as well have carried out classical mechanics on phase space (a symplectic manifold) on which a scalar field, the Hamiltonian, was specified. On the other hand, right at the beginning in quantum mechanics—already in the specification of the states—configuration space plays an essential role. Suppose, for example, that we produced two manifolds whose cotangent bundles were identical as symplectic manifolds. Let this symplectic manifold be the phase space for some system. Then either of our two original manifolds could be regarded as the configuration space of the system. In classical mechanics, this choice makes no difference, for classical mechanics takes place on phase space. But, in quantum mechanics, the difference is crucial, for the space of states already depends essentially on the choice of configuration space. Stated more carefully, one might say that the physical world apparently attaches significance to configuration space itself, but that this significance seems to get lost in the passage to the classical limit.

Essentially, the only structure available on the space of classical states are the smoothness and symplectic structures. What structure is available on the space of quantum states? As it turns out, a great deal more. Firstly, we can add states, if ψ and ψ' are complex-valued densities on C of weight $+\frac{1}{2}$ for which the integral (41) converges, then $\psi+\psi'$ (sum of densities) is another. Furthermore, if ψ is a quantum state, and κ a complex number, then $\kappa\psi$ is another state. Thus, we can add states and multiply them with complex numbers. It is obvious that the space of quantum states thus becomes a complex vector space. Still more

structure comes from (40). Given two states, ψ and ψ', we define their **inner product**, a complex number, by

$$\langle \psi' \,|\, \psi \rangle = \int_C \bar{\psi}' \psi \tag{42}$$

It is clear that the inner product satisfies the following properties:

1. $\langle \psi \,|\, \psi' + \kappa \psi'' \rangle = \langle \psi \,|\, \psi' \rangle + \kappa \langle \psi \,|\, \psi'' \rangle$
2. $\overline{\langle \psi \,|\, \psi' \rangle} = \langle \psi' \,|\, \psi \rangle$
3. $\langle \psi \,|\, \psi \rangle \geq 0$, equality holding when and only when $\psi = 0$

More generally, an **inner product space** is a complex vector space on which an inner product is defined, satisfying conditions (1), (2) and (3) above. Thus, the space of quantum states is an inner product space. An inner product space which is *complete* (in an appropriate sense) is called a **Hilbert space**. The inner product spaces which occur in quantum mechanics are almost never complete. However, there exists a procedure for "completing" an inner product space to obtain a Hilbert space. Thus, one sometimes regards the states of a quantum system as residing in a Hilbert space (*i.e.*, by suitably completing our space of states above). It makes essentially no difference in practice whether one choose to work in the inner product or the Hilbert space. Primarily to avoid having to introduce the technical construction of completing, we shall remain in the inner product space.

It is natural to ask why the very rich structure of the states of an actual system (*i.e.*, the quantum states) become lost in the passage to the classical limit. This is not at all clear (to me) except for the remark that, for some reason, *superposition is lost in the classical limit.*

16. Observables

We now introduce observables in quantum mechanics. It is convenient to begin summarizing some mathematics.

Let $\mathcal{H}$ be an inner product space. A mapping $O : \mathcal{H} \to \mathcal{H}$ which is linear (*i.e.*, which satisfies $O(\psi + \kappa\psi') = O(\psi) + \kappa O(\psi')$ for all elements ψ, ψ' of $\mathcal{H}$ and complex numbers κ) is called an **operator** on $\mathcal{H}$. Actually, it is convenient to slightly broaden this definition. A subspace $\mathcal{D}$ of $\mathcal{H}$ will be said to be **dense** if, given any element ψ of $\mathcal{H}$ and any number $\epsilon > 0$, there is an element ψ' of $\mathcal{D}$ such that $\|\psi - \psi'\| \leq \epsilon$. (Intuitively, a dense subset possesses elements arbitrarily close to any element of $\mathcal{H}$.) We shall also call linear mappings from a dense subspace $\mathcal{D}$ of $\mathcal{H}$ to $\mathcal{H}$ **operators**. (This is a technical point, necessitated by the fact that most of the operators we shall need are only defined, in any reasonable way, on a dense subspace of an inner product space.) An operator will be called **self-adjoint** if, whenever both ψ and ψ' are in the dense subspace on which the operator is defined, $\langle\, \psi' \,|\, O\psi \,\rangle = \langle\, O\psi' \,|\, \psi \,\rangle$.

An **observable** in quantum mechanics is a self-adjoint operator on the inner product space of states.

Of particular interest are observables which have a direct connection with classical observables. Ideally, one would like to formulate a rule for passing from any classical observable (a scalar field on phase space) to a corresponding quantum observable. One has, however, no reason to expect that such a rule will exist—and, apparently, none does. In fact, it is difficult to think, offhand, of any way whatever to pass from a scalar field on phase space to an operator on the space of certain densities on configuration space. From a purely mathematical point of view, it is perhaps surprising that any classical observables have quantum analogues. Perhaps one should ask for less—ask only that the algebraic observables go over to quantum mechanics. But, at this stage, even this appears to be impossible. It turns out that the classical observables which do have simple and unambiguous quantum versions are the configuration and momentum observables.

Let α be a scalar field on configuration space. Then, as we have seen, α may also be regarded as a scalar field on phase space, *i.e.*, as a classical observable. We now define a corresponding quantum observable: $Q(\alpha)\psi = \alpha\psi$. That is to say, $Q(\alpha)$ is the operator on the inner product space of states which, acting on a typical state ψ yields the state obtained by multiplying ψ (a density on C of weight $+\frac{1}{2}$) by α (a density on C of weight 0). The result is another state, so $Q(\alpha)$ is a mapping from states to states. These operators $Q(\alpha)$ will be called **configuration observables**: they are directly analogous to configuration observables in classical mechanics. Note that each $Q(\alpha)$ is self-adjoint: $\langle\, \psi' \,|\, Q(\alpha)\psi \,\rangle = \int \bar{\psi}'(\alpha\psi) = \int (\alpha\bar{\psi}')\psi = \langle\, Q(\alpha)\psi' \,|\, \psi \,\rangle$.

The momentum observables are slightly more subtle. Let ξ^a by a vector field on configuration space, so $\xi^a p_a$ is a classical momentum observable. We wish to write down a corresponding quantum observable. The standard rule for this transition is that of "replacing the classical momentum p_a by $(\frac{\hbar}{i})\nabla_a$". Thus, one might expect that the appropriate quantum operator is that whose action on a state ψ yields $(\frac{\hbar}{i})\xi^a\nabla_a\psi$. However, there are two difficulties with this choice. Firstly, the operator is not self-adjoint. Indeed,

$$\begin{aligned}
\int \bar{\psi}'((\tfrac{\hbar}{i})\xi^a\nabla_a\psi) - \int \big((\tfrac{\hbar}{i})\xi^a\nabla_a\psi'\big)\psi &= (\tfrac{\hbar}{i})\int \big(\bar{\psi}'\xi^a\nabla_a\psi + (\xi^a\nabla_a\bar{\psi}')\psi\big) \\
&= (\tfrac{\hbar}{i})\int \big(\nabla_a(\bar{\psi}'\psi\xi^a) - \bar{\psi}'\psi\nabla_a\xi^a\big) \\
&= -(\tfrac{\hbar}{i})\int \bar{\psi}'\psi\nabla_a\xi^a
\end{aligned}$$

But the right hand side is not in general zero. Even more serious is the fact that

the expression $(\frac{\hbar}{i})\xi^a\nabla_a\psi$ is not independent of the choice of derivative operator. In fact, we have from (40)

$$\xi^a\nabla'_a\psi - \xi^a\nabla_a\psi = \frac{1}{2}\xi^a C^m{}_{am}\psi \neq 0.$$

It is a remarkable fact that a single change in our operator suffices to correct both of these difficulties. For ξ^a, a vector field on configuration space, we define an operator, $P(\xi)$, with the following action on a typical state ψ:

$$P(\xi)\psi = (\tfrac{\hbar}{i})(\xi^a\nabla_a\psi + \tfrac{1}{2}\psi\nabla_a\xi^a) \tag{43}$$

It is immediate from the above that the second term on the right in (43) makes this operator self-adjoint, and, at the same time, produces an operator independent of the choice of derivative operator. (This last property is vital, for we have, on configuration space, no preferred derivative operator.) The quantum observables $P(\xi)$ will be called **momentum observables**. (Note that the usual angular momentum operators in quantum mechanics are also, by our definition, momentum observables.) The quantum momentum observables are, of course, analogous to the classical momentum observables.

There are three fundamental sets of equations satisfied by these configuration and momentum observables. The first, and easiest, are linearity in the argument:

$$Q(\alpha+\beta) = Q(\alpha) + Q(\beta) \tag{44}$$

$$P(\xi+\eta) = P(\xi) + P(\eta) \tag{45}$$

These properties are immediate from the definitions. The second set of equations are two anti-commutator relations. For A and B operators, we define the **anti-commutator** by $\{A,B\} = AB+BA$. We have, immediately from the definition, that $Q(\alpha)Q(\beta) = Q(\alpha\beta)$. Hence,

$$\{Q(\alpha), Q(\beta)\} = 2Q(\alpha\beta) \tag{46}$$

The anti-commutator between configuration and momentum observables also takes a simple form. For ψ a state, we have

$$\begin{aligned}
\{P(\xi),Q(\alpha)\}\psi &= P(\xi)\,(Q(\alpha)\psi) + Q(\alpha)\,(P(\xi)\psi) \\
&= P(\xi)(\alpha\psi) + \alpha(\tfrac{\hbar}{i})(\xi^a\nabla_a\psi + \tfrac{1}{2}\psi\nabla_a\xi^a) \\
&= (\tfrac{\hbar}{i})\left((\xi^a\nabla_a(\alpha\psi) + \tfrac{1}{2}(\alpha\psi)\nabla_a\xi^a) + \alpha(\xi^a\nabla_a\psi + \tfrac{1}{2}\psi\nabla_a\xi^a)\right) \\
&= (\tfrac{\hbar}{i})\,(\xi^a(\nabla_a\alpha)\psi + 2\alpha\xi^a\nabla_a\psi + \alpha\psi\nabla_a\xi^a) \\
&= (\tfrac{\hbar}{i})\,(2\alpha\xi^a\nabla_a\psi + \psi\nabla_a(\alpha\xi^a)) \\
&= 2(\tfrac{\hbar}{i})\left(\alpha\xi^a\nabla_a\psi + \tfrac{1}{2}\psi\nabla_a(\alpha\xi^a)\right) \\
&= 2P(\alpha\xi)\psi
\end{aligned}$$

Hence, since ψ is arbitrary,

$$\{P(\xi), Q(\alpha)\} = 2P(\alpha\xi) \tag{47}$$

Finally, we derive the three (canonical) commutation relations. For A and B operators, we set $[A,B] = AB-BA$. It is immediate from the definition that

$$[Q(\alpha), Q(\beta)] = 0 \tag{48}$$

Furthermore, we have

$$\begin{aligned}
[P(\xi),Q(\alpha)]\psi &= P(\xi)\,(Q(\alpha)\psi) - Q(\alpha)\,(P(\xi)\psi) \\
&= P(\xi)(\alpha\psi) - \alpha(\tfrac{\hbar}{i})(\xi^a\nabla_a\psi + \tfrac{1}{2}\psi\nabla_a\xi^a) \\
&= (\tfrac{\hbar}{i})\left((\xi^a\nabla_a(\alpha\psi) + \tfrac{1}{2}(\alpha\psi)\nabla_a\xi^a) - \alpha(\xi^a\nabla_a\psi + \tfrac{1}{2}\psi\nabla_a\xi^a)\right) \\
&= (\tfrac{\hbar}{i})\,(\xi^a(\nabla_a\alpha)\psi) \\
&= (\tfrac{\hbar}{i})Q(\xi^a\nabla_a\alpha)\psi
\end{aligned}$$

whence

$$[P(\xi), Q(\alpha)] = (\tfrac{\hbar}{i})Q(\xi^a\nabla_a\alpha) \tag{49}$$

Finally, we obtain the commutation relations for the momentum observables:

My n^{th}-generation copy of this page had unreadable sections. So, this page is my attempt to reconstruct the content. I hope that I have maintained the style of the presentation and that I did not introduce any errors. (Please check!)

$$
\begin{aligned}
&[P(\xi), P(\eta)]\psi \\
&= P(\xi)\left[P(\eta)\psi\right] - P(\eta)\left[P(\xi)\psi\right] \\
&= P(\xi)(\tfrac{\hbar}{i})[\eta^b\nabla_b\psi + \tfrac{1}{2}\psi\nabla_b\eta^b] - P(\eta)(\tfrac{\hbar}{i})[\xi^b\nabla_b\psi + \tfrac{1}{2}\psi\nabla_b\xi^b] \\
&= (\tfrac{\hbar}{i})\left[\xi^a\nabla_a((\tfrac{\hbar}{i})[\eta^b\nabla_b\psi + \tfrac{1}{2}\psi\nabla_b\eta^b]) + \tfrac{1}{2}(\tfrac{\hbar}{i})[\eta^b\nabla_b\psi + \tfrac{1}{2}\psi\nabla_b\eta^b](\nabla_a\xi^a)\right] \\
&\quad -(\tfrac{\hbar}{i})\left[\eta^a\nabla_a((\tfrac{\hbar}{i})[\xi^b\nabla_b\psi + \tfrac{1}{2}\psi\nabla_b\xi^b]) + \tfrac{1}{2}(\tfrac{\hbar}{i})[\xi^b\nabla_b\psi + \tfrac{1}{2}\psi\nabla_b\xi^b](\nabla_a\eta^a)\right] \\
&= (\tfrac{\hbar}{i})^2\left[\xi^a\nabla_a(\eta^b\nabla_b\psi) + \tfrac{1}{2}\xi^a\nabla_a(\psi\nabla_b\eta^b) + \tfrac{1}{2}(\eta^b\nabla_b\psi)\nabla_a\xi^a + \tfrac{1}{2}(\tfrac{1}{2}\psi\nabla_b\eta^b)\nabla_a\xi^a\right. \\
&\qquad \left. - \eta^a\nabla_a(\xi^b\nabla_b\psi) - \tfrac{1}{2}\eta^a\nabla_a(\psi\nabla_b\xi^b) - \tfrac{1}{2}(\xi^b\nabla_b\psi)\nabla_a\eta^a - \tfrac{1}{2}(\tfrac{1}{2}\psi\nabla_b\xi^b)\nabla_a\eta^a\right] \\
&= (\tfrac{\hbar}{i})^2\left[(\xi^a\nabla_a\eta^b - \eta^a\nabla_a\xi^b)\nabla_b\psi + (\xi^a\eta^b - \eta^a\xi^b)\nabla_a\nabla_b\psi\right. \\
&\qquad +[\tfrac{1}{2}\psi\xi^a\nabla_a(\nabla_b\eta^b) + \tfrac{1}{2}\xi^a(\nabla_a\psi)\nabla_b\eta^b] + \tfrac{1}{2}(\eta^b\nabla_b\psi)\nabla_a\xi^a \\
&\qquad \left. -[\tfrac{1}{2}\psi\eta^a\nabla_a(\nabla_b\xi^b) + \tfrac{1}{2}\eta^a(\nabla_a\psi)\nabla_b\xi^b] - \tfrac{1}{2}(\xi^b\nabla_b\psi)\nabla_a\eta^a\right] \\
&= (\tfrac{\hbar}{i})^2\left[(\xi^a\nabla_a\eta^b - \eta^a\nabla_a\xi^b)\nabla_b\psi + (\xi^a\eta^b - \eta^a\xi^b)\nabla_a\nabla_b\psi\right. \\
&\qquad \left. +\tfrac{1}{2}\psi[\xi^a\nabla_a\nabla_b\eta^b - \eta^a\nabla_a\nabla_b\xi^b]\right] \\
&= (\tfrac{\hbar}{i})^2\left[(\xi^a\nabla_a\eta^b - \eta^a\nabla_a\xi^b)\nabla_b\psi + (\xi^a\eta^b - \eta^a\xi^b)\nabla_a\nabla_b\psi\right. \\
&\qquad +\tfrac{1}{2}\psi[\xi^a\nabla_a\nabla_b\eta^b - \eta^a\nabla_a\nabla_b\xi^b] - \tfrac{1}{2}\psi[\xi^a\nabla_b\nabla_a\eta^b - \eta^a\nabla_b\nabla_a\xi^b] \\
&\qquad \left. +\tfrac{1}{2}\psi[\xi^a\nabla_b\nabla_a\eta^b - \eta^a\nabla_b\nabla_a\xi^b]\right] \\
&= (\tfrac{\hbar}{i})^2\left[(\xi^a\nabla_a\eta^b - \eta^a\nabla_a\xi^b)\nabla_b\psi + (\xi^a\eta^b - \eta^a\xi^b)\nabla_a\nabla_b\psi\right. \\
&\qquad +\psi[\xi^a\nabla_{[a}\nabla_{b]}\eta^b - \eta^a\nabla_{[a}\nabla_{b]}\xi^b] \\
&\qquad \left. +\tfrac{1}{2}\psi[\nabla_b(\xi^a\nabla_a\eta^b) - (\nabla_b\xi^a)(\nabla_a\eta^b) - \nabla_b(\eta^a\nabla_a\xi^b) + (\nabla_b\eta^a)(\nabla_a\xi^b)]\right] \\
&= (\tfrac{\hbar}{i})^2\left[(\xi^a\nabla_a\eta^b - \eta^a\nabla_a\xi^b)\nabla_b\psi + \tfrac{1}{2}\psi\nabla_b(\xi^a\nabla_a\eta^b - \eta^a\nabla_a\xi^b)\right. \\
&\qquad \left. +(\xi^a\eta^b - \eta^a\xi^b)\nabla_a\nabla_b\psi + \psi[\xi^a\nabla_{[a}\nabla_{b]}\eta^b - \eta^a\nabla_{[a}\nabla_{b]}\xi^b]\right] \\
&= (\tfrac{\hbar}{i})^2\left[(\tfrac{\hbar}{i})^{-1}P(\pounds_\xi\eta)\psi\right. \\
&\qquad \left. +(\xi^a\eta^b - \eta^a\xi^b)\nabla_a\nabla_b\psi + \psi[\xi^a\nabla_{[a}\nabla_{b]}\eta^b - \eta^a\nabla_{[a}\nabla_{b]}\xi^b]\right] \\
&= (\tfrac{\hbar}{i})P(\pounds_\xi\eta)\psi + (\tfrac{\hbar}{i})^2\psi^{-1}\nabla_a\nabla_b(\psi^2\xi^{[a}\eta^{b]})
\end{aligned}
$$

where we have used the identity that

$$
\begin{aligned}
\psi^{-1}\nabla_a\nabla_b(\psi^2\xi^{[a}\eta^{b]}) &= \psi^{-1}[\nabla_{[a}\nabla_{b]}(\psi^2\xi^a\eta^b)] \\
&= \psi^{-1}[\psi^2\nabla_{[a}\nabla_{b]}(\xi^a\eta^b) + \xi^a\eta^b\nabla_{[a}\nabla_{b]}(\psi^2)] \\
&= \psi^{-1}[\psi^2(\xi^a\nabla_{[a}\nabla_{b]}\eta^b + \eta^b\nabla_{[a}\nabla_{b]}\xi^a) + \xi^a\eta^b(2\psi\nabla_{[a}\nabla_{b]}\psi)] \\
&= \psi^{-1}[\psi^2(\xi^a\nabla_{[a}\nabla_{b]}\eta^b - \eta^a\nabla_{[a}\nabla_{b]}\xi^b) + \xi^a\eta^b(2\psi\nabla_{[a}\nabla_{b]}\psi)] \\
&= \psi(\xi^a\nabla_{[a}\nabla_{b]}\eta^b - \eta^a\nabla_{[a}\nabla_{b]}\xi^b) + 2\xi^{[a}\eta^{b]}\nabla_a\nabla_b\psi \\
&= \psi(\xi^a\nabla_{[a}\nabla_{b]}\eta^b - \eta^a\nabla_{[a}\nabla_{b]}\xi^b) + (\xi^a\eta^b - \eta^a\xi^b)\nabla_a\nabla_b\psi
\end{aligned}
$$

The final result

$$[P(\xi), P(\eta)] = (\tfrac{\hbar}{i})P(\pounds_\xi\eta) \tag{50}$$

will follow from the calculation above if we show that the last term on the far right side [*i.e.*, $(\tfrac{\hbar}{i})^2\psi^{-1}\nabla_a\nabla_b(\psi^2\xi^{[a}\eta^{b]})$] vanishes. Setting $F^{ab} = F^{[ab]} = \psi^2\xi^{[a}\eta^{b]}$, we have, for an arbitrary scalar field α

$$\int \alpha\nabla_a\nabla_b F^{ab} = -\int (\nabla_a\alpha)\nabla_b F^{ab} = \int F^{ab}\nabla_a\nabla_b\alpha = \int F^{ab}\nabla_{[a}\nabla_{b]}\alpha = 0$$

where we have integrated twice by parts. Hence, $\nabla_a\nabla_b F^{ab} = 0$, and our result, (50), follows.

Eqns. (48), (49), (50) are called the **canonical commutation relations**. Note that, in each case, the commutator of the observables is $(\frac{\hbar}{i})$ times the Poisson bracket of the corresponding classical observables. It is interesting that the quantum observables manage to encompass, in the commutation relations, two features of the classical observables. Firstly, the commutators all contain a factor of $\hbar$ on the right. Thus, in the classical limit ($\hbar \to 0$), the observables commute, just as do classical observables (where product is outer product). On the other hand, the Poisson bracket of classical observables is also reflected in the commutators.

The fundamental properties of the configuration and momentum observables are given by Eqns. (44)-(50). These operators and properties are incorporated, in a somewhat peculiar way, into the standard formulation of Schrodinger quantum mechanics. Consider, for example, a free particle. One introduces configuration observables Q_x, Q_y, and Q_z, obtained from the scalar fields x, y, and z on Euclidean space (configuration space). These commute with each other. From (44) and (46), configuration observables associated with any scalar field which is a polynomial in x, y, z can be constructed as combinations of the observables Q_x, Q_y, and Q_z. One introduces momentum observables P_x, P_y, and P_z, defined by unit vector fields in the x, y, and z directions, respectively. Using (45) and (47), the momentum observable associated with any vector field in Euclidean space whose Euclidean components are polynomials in x, y, and z can be constructed from Q_x, Q_y, Q_z, P_x, P_y, P_z. Thus, using (44)-(47), one recovers essentially (*i.e.*, insofar as polynomials are "essentially all functions") all of the configuration and momentum observables in terms of just six. On these six observables, the canonical commutation relations are imposed. These, of course, are just (48)-(50) in the special cases. Since all other observables are expressed in terms of these six, the commutation relations on the six suffice to obtain all commutation relations. Our formulae (48)-(50) are just these "general commutation relations for all configuration and momentum observables". In short, instead of many observables, subject to (44)-(50), one uses (44)-(47) to reduce to six observables, on which (48)-(50) are imposed. The advantage of our formulation, in my view, is that one avoids making special choices (of which six observables are to be given special status), and treats everything on a simple, one-shot basis. For examples, the angular momentum commutators for the free particle are special cases of (50).

17. Higher Order Observables

The introduction of configuration and momentum observables works out remarkably well. The definitions are simple and natural, and the resulting operators satisfy every condition one could reasonably ask for. One is led by this success to try to carry over into quantum mechanics classical observables which are quadratic, cubic, *etc.* in momentum. This we attempt, unsuccessfully, in the present section.

First note that it is easy to construct observables which involve second, third, *etc.* derivatives of ψ. For example, let ξ^a be a vector field on configuration space. Then $P(\xi)P(\xi)$ is an observable whose explicit action, from (43), is

$$\begin{aligned} P(\xi)P(\xi)\psi &= (\tfrac{\hbar}{i})^2 \left[\xi^a\xi^b\nabla_a\nabla_b\psi + (\xi^a\nabla_b\xi^b + \xi^b\nabla_b\xi^a)\nabla_a\psi \right. \\ &\qquad \left. +(\tfrac{1}{2}\xi^a\nabla_a(\nabla_b\xi^b) + \tfrac{1}{4}(\nabla_b\xi^b)^2)\psi\right] \end{aligned}$$

More generally, any anti-commutator of observables is an observable. It would be natural to regard $\{P(\xi), P(\eta)\}$ as the quantum observable $T^{ab}p_ap_b$, where $T^{ab} = \xi^{(a}\eta^{b)}$. (That is to say, we suitably generalize (46) and (47).) Unfortunately, a problem arises. Let λ be any scalar field on configuration space C which vanishes nowhere (so λ^{-1} exists). Then it is also true that $T^{ab} = (\lambda^{-1}\xi^{(a})(\lambda\eta^{b)})$. However, using (46)-(50), it is easily checked in general $\{P(\lambda^{-1}\xi), P(\lambda\eta)\} \neq \{P(\xi), P(\eta)\}$. More generally, suppose we are given a quadratic classical observable, $T^{ab}p_ap_b$. One can always write T^{ab} as a sum of symmetrized outer products of vector fields. One would like to call the corresponding sum of anti-commutators of momentum observables the quantum version of the classical observable $T^{ab}p_ap_b$. However, as illustrated above, the resulting quantum observable will in general depend on the decomposition chosen for T^{ab}. In short, we fail, in this way, to associate with classical observables which are quadratic, cubic, *etc.* in momentum corresponding (unique) quantum observables.

Let $T^{ab}p_ap_b$ be a classical observable. We proceed, in a more direct way, to try to associate with this a quantum observable. Clearly, the most general candidate is

$$\alpha(T^{ab})\nabla_a\nabla_b\psi + \beta(\nabla_aT^{ab})\nabla_b\psi + \gamma(\nabla_a\nabla_bT^{ab})\psi \tag{51}$$

where α, β, and γ are real numbers. In order to be acceptable, this candidate must, at least, satisfy two conditions: it must be self-adjoint, and it must be independent of the choice of derivative operator.

My n^{th}-generation copy of this page had unreadable sections. So, this page is my attempt to reconstruct the content. I hope that I have maintained the style of the presentation and that I did not introduce any errors. (Please check!)

We ask what α, β, and γ must be in order that each of these conditions, in turn, be satisfied. Evidently, we have

$$\begin{aligned} &\int \bar\psi'[\alpha(T^{ab})\nabla_a\nabla_b\psi + \beta(\nabla_aT^{ab})\nabla_b\psi + \gamma(\nabla_a\nabla_bT^{ab})\psi] \\ &= \int[-\alpha\nabla_a(\bar\psi'T^{ab})\nabla_b\psi + \beta\bar\psi'(\nabla_aT^{ab})\nabla_b\psi + \gamma\bar\psi'(\nabla_a\nabla_bT^{ab})\psi] \\ &= \int[-\alpha(T^{ab}(\nabla_a\bar\psi')\nabla_b\psi + \bar\psi'(\nabla_aT^{ab})\nabla_b\psi) \\ &\qquad +\beta\bar\psi'(\nabla_aT^{ab})\nabla_b\psi + \gamma\bar\psi'\psi(\nabla_a\nabla_bT^{ab})] \\ &= \int[-\alpha T^{ab}(\nabla_a\bar\psi')\nabla_b\psi + (\beta-\alpha)\bar\psi'(\nabla_aT^{ab})\nabla_b\psi + \gamma\bar\psi'\psi(\nabla_a\nabla_bT^{ab})] \end{aligned}$$

where, in the first step, we have integrated by parts, and, in the third step, we used the symmetry of T^{ab}. Hence, a necessary and sufficient condition that T be self-adjoint is that $\alpha = \beta$. To investigate behavior under change in the choice of derivative operator, we use (40):

$$
\begin{aligned}
&\alpha(T^{ab})\nabla'_a\nabla'_b\psi + \beta(\nabla'_a T^{ab})\nabla'_b\psi + \gamma(\nabla'_a\nabla'_b T^{ab})\psi \\
&= \alpha T^{ab}\nabla'_a(\nabla_b\psi + \tfrac{1}{2}C^m{}_{bm}\psi) \\
&\quad +\beta(\nabla_a T^{ab} - C^a{}_{am}T^{mb} - C^b{}_{am}T^{am})(\nabla_b\psi + \tfrac{1}{2}C^p{}_{bp}\psi) \\
&\quad +\gamma(\nabla'_a(\nabla_b T^{ab} - C^a{}_{bm}T^{mb} - C^b{}_{bm}T^{am}))\psi \\
&= \alpha T^{ab}(\nabla_a\nabla_b\psi + C^m{}_{ab}\nabla_m\psi) \\
&\qquad +\tfrac{1}{2}\alpha T^{ab}(C^m{}_{bm}(\nabla_a\psi + \tfrac{1}{2}C^n{}_{an}\psi) + (\nabla'_a C^m{}_{bm})\psi) \\
&\quad +\beta(\nabla_a T^{ab} - C^a{}_{am}T^{mb} - C^b{}_{am}T^{am})\nabla_b\psi \\
&\qquad +\tfrac{1}{2}\beta(\nabla_a T^{ab} - C^a{}_{am}T^{mb} - C^b{}_{am}T^{am})C^p{}_{bp}\psi \\
&\quad +\gamma(\nabla_a\nabla_b T^{ab} - C^a{}_{am}\nabla_b T^{mb})\psi + \gamma(-\nabla'_a(C^a{}_{bm}T^{mb} + C^b{}_{bm}T^{am}))\psi \\
&= \alpha T^{ab}\nabla_a\nabla_b\psi + \beta(\nabla_a T^{ab})\nabla_b\psi + \gamma(\nabla_a\nabla_b T^{ab})\psi \\
&\quad +\alpha T^{ab}C^m{}_{ab}\nabla_m\psi + \tfrac{1}{2}\alpha T^{ab}C^m{}_{bm}\nabla_a\psi \\
&\qquad +(\tfrac{1}{2})^2\alpha T^{ab}C^m{}_{bm}C^n{}_{an}\psi + \tfrac{1}{2}\alpha T^{ab}(\nabla'_a C^m{}_{bm})\psi \\
&\quad -\beta C^a{}_{am}T^{mb}\nabla_b\psi - \beta C^b{}_{am}T^{am}\nabla_b\psi \\
&\qquad +\tfrac{1}{2}\beta(\nabla_a T^{ab} - C^a{}_{am}T^{mb} - C^b{}_{am}T^{am})C^p{}_{bp}\psi \\
&\quad -\gamma(C^a{}_{am}\nabla_b T^{mb} + \nabla'_a(C^a{}_{bm}T^{mb} + C^b{}_{bm}T^{am}))\psi \\
&= \alpha T^{ab}\nabla_a\nabla_b\psi + \beta(\nabla_a T^{ab})\nabla_b\psi + \gamma(\nabla_a\nabla_b T^{ab})\psi \\
&\quad +(\nabla_a\psi)((\tfrac{1}{2}\alpha - \beta)T^{ab}C^m{}_{bm} + (\alpha - \beta)T^{mb}C^a{}_{mb}) \\
&\quad +(\psi)\left(\tfrac{1}{2}\alpha T^{ab}(\tfrac{1}{2}C^m{}_{bm}C^n{}_{an} + \nabla'_a C^m{}_{bm})\right. \\
&\qquad +\tfrac{1}{2}\beta(\nabla_a T^{ab} - C^a{}_{am}T^{mb} - C^b{}_{am}T^{am})C^p{}_{bp} \\
&\qquad \left. -\gamma(C^a{}_{am}\nabla_b T^{mb} + \nabla'_a(C^a{}_{bm}T^{mb} + C^b{}_{bm}T^{am}))\right)
\end{aligned}
$$

My n^{th}-generation copy of this page had unreadable sections. So, this page is my attempt to reconstruct the content. I hope that I have maintained the style of the presentation and that I did not introduce any errors. (Please check!)

It is immediate from this expression that our operator is independent of the choice of derivative operator precisely when $\alpha = \beta = \gamma = 0$, *i.e.*, when and only when we have the zero operator. This failure is, of course, just a more sophisticated version of our earlier failure.

Thus, we reach the conclusion that it is only the configuration and momentum observables which go over naturally from classical to quantum mechanics. But, as we remarked earlier, what is perhaps surprising is, not that some classical observables have no quantum analogs, but that any at all do. Nevertheless, one is led to ask why the cut-off occurs between linearity and quadratic-ness in momentum. (Mathematically, it is because different derivative operators only start to become really different after the second derivative.) There is, however, a more direct reason.

Recall that every classical observable generates an infinitesimal canonical transformation. Let us call a canonical transformation "configuration space preserving" if it has the following property: *two points of phase space associated with the same point of configuration space (e.g., (q, p_a) and (q, p'_a)) are taken, by the canonical transformation, to points of phase space which again have this property.* Note that a classical observable is a configuration observable if and only if it assumes the same value on two points of phase space associated with the same point of configuration space. Thus, a canonical transformation is configuration space preserving if and only if it takes configuration observables to configuration observables. The infinitesimal version of this statement is this: *the infinitesimal canonical transformation generated by an observable is to be regarded as configuration space preserving when and only when the Poisson bracket of that observable with any configuration observable is a configuration observable.* But now (37) implies (immediately) that the only observables which generate infinitesimal canonical transformations which are configuration space

preserving are the configuration and momentum observables. Thus, these classical observables play a special role.

What is so special about things which are configuration space preserving? Recall that quantum mechanics, as opposed to classical mechanics, requires, not just phase space, but also configuration space itself. Classical observables which "disrupt" configuration space (*i.e.*, whose infinitesimal canonical transformations are not configuration space preserving) might be expected to be awkward in quantum mechanics. Indeed, as we have found, they are awkward; they do not seem to carry over from classical to quantum mechanics.

18. The Hamiltonian. Dynamics

We now wish to describe the passage of our quantum system, with time, from one state to another. In classical mechanics, this description is carried out by introducing a certain observable H, the Hamiltonian. This H then defines the dynamical trajectories, which describe the dynamics. Thus, the most obvious way to proceed is to first to try to replace the classical observable H by some sort of analogous quantum observable.

We are confronted immediately by a serious problem. We have seen in Sect. 17 that it is only the configuration and momentum observables which go over directly from classical to quantum mechanics. However, the classical Hamiltonian is almost never a configuration or momentum observable. In fact, it is almost always the case that the Hamiltonian contains quadratic terms in the momenta, *i.e.*, that H takes the form

$$H = g^{ab}p_a p_b + A^a p_a + V \tag{52}$$

Such a classical observable, by Sect. 17, does not go over in the obvious way to a quantum observable.

Fortunately, a remarkably simple result intervenes to save the situation. It is this: *On a manifold M, let g^{ab} be a symmetric (i.e., $g^{ab} = g^{(ab)}$) tensor field which is invertible (i.e., for which there exists a field g_{ab} such that $g^{am}g_{bm} = \delta^a{}_b$). Then, there exists one and only one derivative operator ∇_a on M such that $\nabla_a g^{bc} = 0$.* Proof: Let ∇'_a be any derivative operator on M. We then have, from (6),

$$\nabla_a g^{bc} = \nabla'_a g^{bc} + C^b{}_{am}g^{mc} + C^c{}_{am}g^{bm} \tag{53}$$

We must first show that there is one and only one tensor field $C^a{}_{bc}$ on M such that the right hand side of (53) vanishes. Indeed, there is one, namely,

$$C^a{}_{bc} = -g_{bm}\nabla_c g^{am} - g_{cm}\nabla_b g^{am} + g^{am}g_{bp}g_{cq}\nabla_m g^{pq} \tag{54}$$

as can be checked by substituting (54) into the right hand side of (53). Suppose both $C^a{}_{bc}$ and $\tilde{C}^a{}_{bc}$ caused the right side of (53) to vanish. Then, setting $K_{abc} = g_{am}(C^m{}_{bc} - \tilde{C}^m{}_{bc})$ we have $K_{abc} = K_{a(bc)}$, and, from the vanishing of the right side of (53), $K_{abc} = -K_{bac}$. Using alternatively the first and second of these facts, we have

$$K_{abc} = K_{acb} = -K_{cab} = -K_{cba} = K_{bca} = K_{bac} = -K_{abc}$$

whence $K_{abc} = 0$, whence $C^a{}_{bc} = \tilde{C}^a{}_{bc}$. This completes the proof.

The idea is to use this result to convert the classical Hamiltonian (52) into a quantum observable. In order to do this, we must assume that the g^{ab} in (52) is invertible. (This condition is certainly satisfied for all the usual systems one treats in quantum mechanics.) Thus, if the classical Hamiltonian contains a term quadratic in momentum (together, possibly, with terms linear in momentum and independent of momentum), and if the coefficient of that term, g^{ab}, is invertible, we can introduce the corresponding quantum operator

$$H\psi = (\tfrac{\hbar}{i})^2 g^{ab}\nabla_a\nabla_b\psi + (\tfrac{\hbar}{i})(A^a\nabla_a\psi + \tfrac{1}{2}\psi\nabla_a A^a) + V\psi \tag{55}$$

where ∇_a in (55) is the unique derivative operator defined by g^{ab} (*i.e.*, such that $\nabla_c g^{ab} = 0$). We needn't be concerned with whether or not the right side of (55) is independent of derivative operator (it isn't), because the classical Hamiltonian prefers one derivative operator. Note that since $\nabla_a g^{bc} = 0$, the first term on the right in (55) is the same as $(\frac{\hbar}{i})^2\nabla_a(g^{ab}\nabla_b\psi)$, and the same as $(\frac{\hbar}{i})^2\nabla_a\nabla_b(g^{ab}\psi)$. (In more conventional language, with this choice of derivative operator, there is no factor ordering problem.) That is to say, if T^{ab} in (51) is replaced by g^{ab}, the last two terms on the right vanish. It is now immediate that from the discussion following (51) that the operator (55) is self-adjoint.

Thus, if the classical Hamiltonian is of the form (52), with g^{ab} invertible, we can introduce a unique corresponding quantum Hamiltonian (55). As far as I am aware, it is only when the classical Hamiltonian satisfies these conditions that one can "quantize" a classical system. Perhaps one should take the point of view that systems (*i.e.*, quantum theory) operate only within this regime. What is then somewhat mysterious is why it should be that, when one passes to the classical limit, the description has a very natural generalization (*i.e.*, to Hamiltonians other than (52)) which is completely lost within quantum mechanics itself. Alternatively, one could take the following view. In quantum mechanics, there is a Hamiltonian observable, which is of finite differential order. A system described by such a Hamiltonian can, in the classical limit, be described by a classical Hamiltonian algebraic in the momenta. It just happens that it is only when the original quantum Hamiltonian was of second differential order that one can pass uniquely back from classical to the quantum description. But, a priori, one might not have expected to be able to pass from the classical description, uniquely, to the quantum description in any case. It just happens that, in the quadratic case, no essential information gets lost in the classical limit. (Nature, however, tries to confuse use by making most quantum Hamiltonians of second differentiable order, so we get used to the idea that we should be able to recover the quantum description from its classical limit.)

A useful simplification becomes available in the presence of the Hamiltonian (55). We claim: *there are just two alternating tensors (differing only in sign) satisfying:*

$$g_{a_1b_1}\cdots g_{a_nb_n}\epsilon^{a_1\ldots a_n}\epsilon^{b_1\ldots b_n} = n!$$

That the left side of this equation is nonzero for any nonzero alternating tensor follows from the fact that g_{ab} is invertible. (This is the tensor statement of the fact from matrix algebra that an invertible matrix has nonzero determinant.) But any multiple of an alternating tensor is an alternating tensor: hence, there exists an alternating tensor satisfying the above. That there are just two, differing in sign, follows from the fact that, if ϵ and $\alpha\epsilon$ satisfy the above, then $\alpha^2 = 1$, whence $\alpha = \pm 1$. Thus, our claim is proved. We can use this preferred alternating tensor (pick one) to reduce density fields to tensor fields. Thus, instead of a density ψ (a mapping from alternating tensors to scalars), we can evaluate this ψ on our preferred alternating tensor $\epsilon^{a_1\ldots a_n}$, to obtain a complex scalar field $\psi(\epsilon)$. Similarly, all tensor densities can be "evaluated on ϵ" to yield tensor fields. Note, furthermore, that, taking the derivative of the equation above, using the derivative operator defined by g^{ab}, we obtain $\nabla_a\epsilon^{a_1\ldots a_n} = 0$. Hence, $\lambda_a = 0$ in (39). That is to say, if we use the derivative operator defined by g^{ab}, and evaluate tensor densities on the preferred alternating tensor to obtain tensor fields, we can replace derivatives of such densities by derivatives of the corresponding tensor fields. (In other words, the operations *"evaluate on the preferred alternating tensor"* and *"take the derivative by the preferred derivative operator"* commute.)

Our final task is to write down the equation describing the dynamics of our system. At each time t, the system is to be in some state, $\psi(t)$, where $\psi(t)$ is (by the previous paragraph) a complex scalar field on configuration space C, for each t. The dynamical equation, called **Schrodinger's equation**, is

$$-(\tfrac{\hbar}{i})\frac{d}{dt}\psi = H\psi \tag{56}$$

This gives the time rate of change of ψ in terms of ψ at one instant of time. Hence, (56) determines $\psi(t)$ for all t, given $\psi(t_0)$ at one fixed time, t_0.

19. Higher Order Observables—Revisited

We have seen in Sect. 17 that classical observables which are of order quadratic or higher in momentum do not lead to quantum observables. The difficulty was this: *no expression could be found for the quantum observable which was independent of the choice of derivative operator.* But, in Sect. 18, we observed that the presence of a (classical) Hamiltonian quadratic in momentum leads, among other things, to a natural, unique derivative operator. Thus released from what was the primary difficulty in Sect. 17, we return to the question of that section.

We begin with the quadratic case. Let $T^{ab}p_ap_b$ be a classical observable. Then, using the unique derivative operator defined by the Hamiltonian, we have the following corresponding quantum observable

$$\tfrac{1}{2}(\tfrac{\hbar}{i})^2[T^{ab}\nabla_a\nabla_b + (\nabla_a\nabla_bT^{ab}) + \gamma(\nabla_bT^{ab})\nabla_a] \tag{57}$$

where γ is any real number. This operator is self-adjoint for all γ, but, of course, is independent of the choice of ∇_a for no γ. In short, one has, in the presence of a preferred derivative operator, a one-parameter family of candidates for the quantum version of the observable $T^{ab}p_ap_b$. It is clear that this situation continues into higher-order observables. A classical observable of quadratic or higher order in momentum leads, in the presence of a preferred derivative operator, to a family of "corresponding" quantum observables depending on some arbitrary constant parameters.

Thus, one before had no quantum versions of higher-order classical observables —one now has many. It is natural, in such a situation, to attempt to impose some additional conditions on our quantum observables in order to obtain uniqueness. There is a natural such condition: one could require that $(\frac{\hbar}{i})$ times Poisson bracket of classical observables correspond to the commutator of the corresponding quantum observables. Consider, for example, the classical observables $T^{ab}p_ap_b$ and ξ^ap_a, with Poisson bracket $(2T^{m(a}\nabla_m\xi^{b)} - \xi^m\nabla_mT^{ab})p_ap_b$ (Sect. 12). The corresponding quantum observables are to be (57) and $(\frac{\hbar}{i})(\xi^a\nabla_a + \frac{1}{2}\nabla_a\xi^a)$. We ask that the commutator of these two observables be $(\frac{\hbar}{i})$ times the operator obtained by replacing T^{ab} in (57) by $(2T^{m(a}\nabla_m\xi^{b)} - \xi^m\nabla_mT^{ab})$. The problem is to choose the constant γ in (57) so that this will be the case. The result of this straightforward and tedious calculation is that no choice of γ does the job. *There is, apparently, no way to assign quantum versions to algebraic classical observables, in the presence of a preferred derivative operator, so that commutators correspond to Poisson brackets.*

Having made this observation, we next observe that it is the answer to what is perhaps a somewhat artificial question. This question of the relation between classical and quantum observables can be reformulated as follows. On a manifold C, a **differential operator** consists of $(\nabla_a, T^{a_1\ldots a_s}, T^{a_1\ldots a_{s-1}}, \ldots, T^a, T)$, where ∇_a is a derivative operator on C, and the T's are symmetric tensor fields on C. The integer s (if $T^{a_1\ldots a_s}$ is nonzero) is the order of the differential operator. We can regard differential operators as acting on scalar densities ψ of weight $+\frac{1}{2}$ as follows:

$$(\tfrac{\hbar}{i})^s[T^{a_1\ldots a_s}\nabla_{a_1}\ldots\nabla_{a_s} + \cdots + T^a\nabla_a + T]\psi \tag{58}$$

Now suppose, in (58), that we rewrite this expression in terms of some other derivative operator ∇'_a, related to ∇_a via $C^a{}_{bc}$ (Eqn. 40). Then (58) becomes

$$(\tfrac{\hbar}{i})^s[T'^{a_1\ldots a_s}\nabla'_{a_1}\ldots\nabla'_{a_s} + \cdots + T'^a\nabla'_a + T']\psi \tag{59}$$

where each $T'^{a\ldots c}$ is expressed in terms of the T's, $C^a{}_{bc}$, and ∇_a. But (59) is to be regarded as the action of the differential operator $(\nabla'_a, T'^{a_1\ldots a_s}, T'^{a_1\ldots a_{s-1}}, \ldots, T'^a, T')$ on ψ. We regard these two differential operators as equivalent. (More precisely, a differential operator is an equivalence class of objects $(\nabla_a, T^{a_1\ldots a_s},$

$T^{a_1\cdots a_{s-1}}, \ldots, T^a, T$), where two such objects are equivalent if their actions, (58), on any density ψ of weight $+\frac{1}{2}$, are identical.)

The quantum observables include self-adjoint differential operators. The classical limit of the quantum observable (58) is, of course, $T^{a_1\cdots a_s} p_{a_1} \cdots p_{a_s}$. Thus, the quantum theory itself has no difficulty in any case with observables. There are many of them lying around, including, presumably, a Hamiltonian observable (possibly quadratic, but also possibly some other order in momentum). We, however, insist on asking about the possibilities for recovering various quantum observables from their classical limit. The situation then becomes quite complicated. Nature, however, doesn't care whether or not quantum observables are recoverable from classical limits.

20. Heisenberg Formulation

What we have been discussing so far is called the Schrodinger formulation of quantum mechanics. There is a second, more natural but essentially equivalent mathematically, formulation called the *Heisenberg formulation of quantum mechanics*. We now introduce it.

It is convenient, in order to illustrate what is involved here, to first return to classical mechanics. Recall that the states in classical mechanics were the points of phase space Γ_C, while the observables were scalar fields on phase space. That is to say, the observables were real-valued functions on the space of states. We may call this the *Schrodinger formulation of classical mechanics.* By contrast, the *Heisenberg formulation of classical mechanics* would be this. The states would be the dynamical trajectories on phase space, *i.e.*, the curves $\gamma(t)$ on phase space, labeled by the time t, whose tangent at each point is the Hamiltonian vector field evaluated at that point. Thus, in the Schrodinger formulation of classical mechanics, the system possesses a state (a point of phase space) at each instant of time: as time marches on, the state changes. In the Heisenberg formulation, a system has one state (a dynamical trajectory), not at each instant of time, but once and for all. Although the system evolves, it always retains the same state, in the Heisenberg formulation. One summarizes the state of affairs by saying that, in the Schrodinger formulation, the state of the system depends on time, while in the Heisenberg formulation, the state is independent of time.

In the Schrodinger formulation of classical mechanics, an observable is a real-valued function on the space of (Schrodinger) states. Similarly, in the Heisenberg formulation, an observable is a real-valued function on the space of (Heisenberg) states. One could, of course, make the space of Heisenberg states into a manifold, so the Heisenberg observables would functions on this manifold. It is convenient, however, to proceed in a slightly different way. Let A be a Schrodinger observable, so A is a scalar field on phase space. We introduce a Heisenberg observable, $O(A, t_0)$, which depends both on A and a choice of time, t_0. This observable is defined as follows: it assigns to the Heisenberg state $\gamma(t)$ (a dynamical trajectory) the number $A(\gamma(t_0))$. That is to say, $O(A, t_0)$ assigns to the Heisenberg state $\gamma(t)$ the number assigned by the Schrodinger observable A to the Schrodinger state $\gamma(t_0)$. Thus, a single Schrodinger observable leads to a one-parameter family $O(A, t)$ (labeled by the parameter t) of Heisenberg observables.

To summarize, a Heisenberg state is a certain one-parameter (labeled by t) family of Schrodinger states. Whereas the Schrodinger state of a system changes with time (as the system evolves), the Heisenberg state does not change with time. In either case, the observables are real-valued functions on the space of states. Each Schrodinger observable leads to a one-parameter family (labeled by t) of Heisenberg observables.

Since the Heisenberg observables which arise from Schrodinger observables depend on the parameter t, it is natural to ask for their rate of change with respect to t. Intuitively, the difference between $O(A, t_0 + \Delta t_0)$ and $O(A, t_0)$, acting on a state $\gamma(t)$, is the difference between A evaluated at the point $\gamma(t_0 + \Delta t_0)$ and the point $\gamma(t_0)$ of phase space. But the Hamiltonian vector field is the tangent vector to a dynamical trajectory. It is immediate, therefore, from (17) that

$$\frac{d}{dt}O(A, t) = O([H, A], t) \tag{60}$$

This equation describes the time-dependence of Heisenberg observables which arise from Schrodinger observables. It also shows, for example, that $O(H, t)$ is independent of t.

Of course, the "reformulation" of classical mechanics above adds practically nothing to the content if classical mechanics. It is of interest because it provides

an almost perfect analogy for the Heisenberg reformulation of Schrodinger quantum mechanics (which, again, adds practically nothing to quantum mechanics.)

In the Schrodinger formulation of quantum mechanics, a state is a density ψ on configuration space C, of weight $+\frac{1}{2}$, which is square-integrable. In the Heisenberg formulation, a state is a solution, $\psi(t)$, of Schrodinger's equation, (56). That is to say, a Heisenberg state is a certain one-parameter family (labeled by t) of Schrodinger states. In the Schrodinger formulation, a system possesses a state at each instant of time, the change in that state with time describing the evolution of the system with time. In the Heisenberg formulation, the system possesses just one state once and for all, the "time evolution" of the system being inherent in the state. Thus, the Schrodinger state describing a system is time-dependent, while the Heisenberg state is time-independent.

In the Schrodinger formulation, an observable is a self-adjoint operator on the space of (Schrodinger) states. In the Heisenberg formulation, an observable is a self-adjoint operator on the space of (Heisenberg) states.

Let A be a Schrodinger observable, so A is a self-adjoint operator on the space of Schrodinger states. We associate with this A a one-parameter family, $O(A, t_0)$, (labeled by time t_0) of Heisenberg observables. The definition is this: if $\psi(t)$ is a Heisenberg state, $O(A, t_0)$, acting on this $\psi(t)$, is that Heisenberg state $\psi'(t)$ such that $\psi'(t_0) = A\psi(t_0)$. In other words, the action of $O(A, t_0)$ is this. Given a Heisenberg state $\psi(t)$, we "freeze" this state at t_0 to obtain a Schrodinger state, $\psi(t_0)$. We then act on this Schrodinger state with A to obtain another Schrodinger state. We then evolve this Schrodinger state to other times (*i.e.*, other than t_0) using the Schrodinger equation. The result is another Heisenberg state. Note that we are forced to a definition of this type. We could not simply take the Heisenberg state $\psi(t)$, and consider $A\psi(t)$ (*i.e.*, the one-parameter family of states obtained by acting on each of the one-parameter family of states $\psi(t)$ with A), for $A\psi(t)$ would not in general satisfy the Schrodinger equation, *i.e.*, would not in general be a Heisenberg state.

To summarize, a Heisenberg state is a certain one-parameter family (labeled by t) of Schrodinger states. An evolving system is described, in the Schrodinger formulation, by a time-dependent Schrodinger state, and, in the Heisenberg formulation, by a single (time-independent) Heisenberg state. Observables, in either case, are self-adjoint operators on the space of states. Every Schrodinger observable defines a one-parameter family of Heisenberg observables.

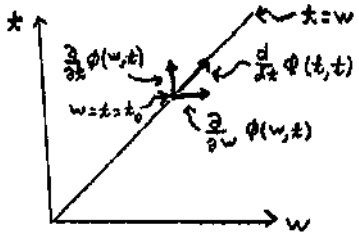
Finally, we wish to derive the quantum equation analogous to (60). Fix a Schrodinger observable A, and a Heisenberg state $\psi(t)$ (a one-parameter family of Schrodinger states). For each real number w, $O(A, w)$, acting on the Heisenberg state $\psi(t)$, yields another Heisenberg state $\phi_w(t)$. We can just as well write this as $\phi(w, t)$, a two-parameter family of Schrodinger states. Since $\phi_w(t)$ is a Heisenberg state for each w, we have

$$-(\tfrac{\hbar}{i})\frac{\partial}{\partial t}\phi(w,t) = H\phi(w,t) \tag{61}$$

for each w. By definition of $O(A, w)$, we have

$$\phi(t,t) = A\psi(t) \tag{62}$$

for each t. Clearly, (61) and (62) define the two-parameter family of Schrodinger states uniquely in terms of the one-parameter family $\psi(t)$. But, from the diagram below,



My n^{th}-generation copy of this page had unreadable sections. So, this page is my attempt to reconstruct the content. I hope that I have maintained the style of the presentation and that I did not introduce any errors. (Please check!)

we see that

$$\begin{aligned}
\frac{\partial}{\partial w}\phi(w,t)\Big|_{w=t=t_0} &= \frac{d}{dt}\phi(t,t)\Big|_{w=t=t_0} - \frac{\partial}{\partial t}\phi(w,t)\Big|_{w=t=t_0} \\
&= \frac{d}{dt}[A\psi(t)]\Big|_{w=t=t_0} - [-(\tfrac{\hbar}{i})^{-1}H\phi(w,t)]\Big|_{w=t=t_0} \\
&= A\frac{d}{dt}\psi(t)\Big|_{w=t=t_0} - [-(\tfrac{\hbar}{i})^{-1}H\phi(w,t)]\Big|_{w=t=t_0} \\
&= -(\tfrac{\hbar}{i})^{-1}AH\psi(t)\big|_{w=t=t_0} + (\tfrac{\hbar}{i})^{-1}HA\psi(t)\big|_{w=t=t_0}
\end{aligned}$$

where, in the last step, we have used (61), (62), and Schrodinger's equation on $\psi(t)$. It is immediate from this equation that

$$\frac{d}{dt}O(A,t) = (\tfrac{\hbar}{i})^{-1}O([H,A],t) \tag{63}$$

This is the desired formula. It gives the rate of change with time of the time-dependent Heisenberg observable arising from any Schrodinger observable A. Note that it implies, in particular, that the Heisenberg observable $O(H,t)$ is independent of t.

It is of interest to work out the right side of (61) for our (Schrodinger) configuration and momentum observables. We take, for our Hamiltonian, (55). Let α be a scalar field on configuration space, and $Q(\alpha)$ the corresponding (Schrodinger) observable. Then

$$\begin{aligned}
&(HQ(\alpha) - Q(\alpha)H)\psi \\
&= [(\tfrac{\hbar}{i})^2 g^{ab}\nabla_a\nabla_b + (\tfrac{\hbar}{i})(A^a\nabla_a + \tfrac{1}{2}\nabla_a A^a) + V](\alpha\psi) \\
&\qquad -\alpha[(\tfrac{\hbar}{i})^2 g^{ab}\nabla_a\nabla_b + (\tfrac{\hbar}{i})(A^a\nabla_a + \tfrac{1}{2}\nabla_a A^a) + V]\psi \\
&= (\tfrac{\hbar}{i})^2[\psi g^{ab}\nabla_a\nabla_b\alpha + 2g^{ab}\nabla_a\alpha\nabla_b\psi] + (\tfrac{\hbar}{i})[\psi A^a\nabla_a\alpha] \\
&= (\tfrac{\hbar}{i})[(\tfrac{\hbar}{i})2(g^{ab}\nabla_a\alpha\nabla_b\psi + \tfrac{1}{2}g^{ab}\nabla_a\nabla_b\alpha)\psi] + (\tfrac{\hbar}{i})[A^a\nabla_a\alpha]\psi \\
&= (\tfrac{\hbar}{i})[2(\tfrac{\hbar}{i})\left((g^{ab}\nabla_a\alpha)\nabla_b\psi + \tfrac{1}{2}g^{ab}\nabla_b\nabla_a\alpha\right)\psi] + (\tfrac{\hbar}{i})[A^a\nabla_a\alpha]\psi \\
&= (\tfrac{\hbar}{i})[2(\tfrac{\hbar}{i})\left((g^{ab}\nabla_a\alpha)\nabla_b\psi + \tfrac{1}{2}\nabla_b(g^{ab}\nabla_a\alpha)\right)\psi] + (\tfrac{\hbar}{i})[A^a\nabla_a\alpha]\psi \\
&= (\tfrac{\hbar}{i})[2(\tfrac{\hbar}{i})\left((g^{ab}\nabla_a\alpha)\nabla_b\psi + \tfrac{1}{2}\psi\nabla_b(g^{ab}\nabla_a\alpha)\right)] + (\tfrac{\hbar}{i})[A^a\nabla_a\alpha]\psi \\
&= (\tfrac{\hbar}{i})[P(2g^{ab}\nabla_a\alpha)\psi] + (\tfrac{\hbar}{i})[Q(A^a\nabla_a\alpha)]\psi \\
&= (\tfrac{\hbar}{i})[P(2g^{ab}\nabla_a\alpha) + Q(A^a\nabla_a\alpha)]\psi
\end{aligned}$$

Note that this reduces, for the case of a particle in Euclidean space, to the usual Heisenberg equation of motion.

My n^{th}-generation copy of this page had unreadable sections. So, this page is my attempt to reconstruct the content. I hope that I have maintained the style of the presentation and that I did not introduce any errors. (Please check!)

The corresponding equation for the momentum operator, $P(\xi)$, is more complicated:

$$
\begin{aligned}
&(HP(\xi)-P(\xi)H)\psi \\
&= [(\tfrac{\hbar}{i})^2 g^{ab}\nabla_a\nabla_b + (\tfrac{\hbar}{i})(A^a\nabla_a + \tfrac{1}{2}\nabla_a A^a) + V][[(\tfrac{\hbar}{i})\xi^c\nabla_c + \tfrac{1}{2}\nabla_c\xi^c]\psi) \\
&\qquad -[(\tfrac{\hbar}{i})\xi^c\nabla_c + \tfrac{1}{2}\nabla_c\xi^c][(\tfrac{\hbar}{i})^2 g^{ab}\nabla_a\nabla_b + (\tfrac{\hbar}{i})(A^a\nabla_a + \tfrac{1}{2}\nabla_a A^a) + V]\psi \\
&= (\tfrac{\hbar}{i})^3\left[g^{ab}\nabla_a\nabla_b(\xi^c\nabla_c\psi + \tfrac{1}{2}(\nabla_c\xi^c)\psi)\right. \\
&\qquad \left. -\xi^c\nabla_c(g^{ab}\nabla_a\nabla_b\psi) - \tfrac{1}{2}(\nabla_c\xi^c)(g^{ab}\nabla_a\nabla_b\psi)\right] \\
&\quad +(\tfrac{\hbar}{i})^2\left[(A^a\nabla_a + \tfrac{1}{2}(\nabla_a A^a))(\xi^c\nabla_c\psi + \tfrac{1}{2}(\nabla_c\xi^c)\psi)\right. \\
&\qquad \left. -(\xi^c\nabla_c + \tfrac{1}{2}(\nabla_c\xi^c))(A^a\nabla_a\psi + \tfrac{1}{2}(\nabla_a A^a)\psi)\right] \\
&\quad +(\tfrac{\hbar}{i})\left[V(\xi^c\nabla_c\psi + \tfrac{1}{2}(\nabla_c\xi^c)\psi) - (\xi^c\nabla_c + \tfrac{1}{2}(\nabla_c\xi^c))(V\psi)\right] \\
&= (\tfrac{\hbar}{i})^3\left[[g^{ab}\nabla_a\nabla_b\xi^c]\nabla_c\psi + [2g^{ab}\nabla_{(a}\xi^c]\nabla_{b)}\nabla_c\psi + [\xi^c g^{ab}]\nabla_a\nabla_b\nabla_c\psi\right. \\
&\qquad +\tfrac{1}{2}[g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi + \tfrac{1}{2}[2g^{ab}\nabla_{(a}(\nabla_c\xi^c)]\nabla_{b)}\psi \\
&\qquad\quad +\tfrac{1}{2}[g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi + \tfrac{1}{2}[(\nabla_c\xi^c)g^{ab}]\nabla_a\nabla_b\psi \\
&\qquad \left. -\xi^c\nabla_c(g^{ab}\nabla_a\nabla_b\psi) - \tfrac{1}{2}(\nabla_c\xi^c)(g^{ab}\nabla_a\nabla_b\psi)\right] \\
&\quad +(\tfrac{\hbar}{i})^2\left[(\tfrac{\hbar}{i})^{-2}(P(A)P(\xi) - P(\xi)P(A))\psi\right] \\
&\quad +(\tfrac{\hbar}{i})\left[(\tfrac{\hbar}{i})^{-1}(Q(V)P(\xi) - P(\xi)Q(V))\psi\right] \\
&= (\tfrac{\hbar}{i})^3\left[[g^{ab}\nabla_a\nabla_b\xi^c]\nabla_c\psi + [2g^{ab}\nabla_{(a}\xi^c]\nabla_{b)}\nabla_c\psi + [\xi^c g^{ab}]\nabla_a\nabla_b\nabla_c\psi\right. \\
&\qquad \left. +2\tfrac{1}{2}[g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi + \tfrac{1}{2}[2g^{ab}\nabla_{(a}(\nabla_c\xi^c)]\nabla_{b)}\psi - \xi^c\nabla_c(g^{ab}\nabla_a\nabla_b\psi)\right] \\
&\quad +(\tfrac{\hbar}{i})^2\left[(\tfrac{\hbar}{i})^{-2}[P(A),P(\xi)]\psi\right] + (\tfrac{\hbar}{i})\left[(\tfrac{\hbar}{i})^{-1}[Q(V),P(\xi)]\psi\right] \\
&= (\tfrac{\hbar}{i})^3\left[[\xi^c g^{ab}](\nabla_a\nabla_b\nabla_c\psi - \nabla_c\nabla_a\nabla_b\psi) + [2g^{ab}\nabla_{(a}\xi^c]\nabla_{b)}\nabla_c\psi\right. \\
&\qquad \left. +[g^{ab}\nabla_a\nabla_b\xi^c]\nabla_c\psi + [g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi + [g^{ab}\nabla_{(a}(\nabla_c\xi^c)]\nabla_{b)}\psi\right] \\
&\quad +(\tfrac{\hbar}{i})^2\left[(\tfrac{\hbar}{i})^{-2}(-(\tfrac{\hbar}{i})P(\pounds_\xi A)\psi)\right] + (\tfrac{\hbar}{i})\left[(\tfrac{\hbar}{i})^{-1}(-(\tfrac{\hbar}{i})Q(\xi^a\nabla_a V)\psi)\right] \\
&= (\tfrac{\hbar}{i})^3\left[[\xi^c g^{ab}](\nabla_a\nabla_c\nabla_b\psi - \nabla_c\nabla_a\nabla_b\psi) + [2g^{ab}\nabla_{(a}\xi^c]\nabla_{b)}\nabla_c\psi\right. \\
&\qquad \left. +[g^{ab}\nabla_a\nabla_b\xi^c]\nabla_c\psi + [g^{ab}\nabla_{(a}(\nabla_c\xi^c)]\nabla_{b)}\psi + [g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi\right] \\
&\quad -(\tfrac{\hbar}{i})\left[P(\pounds_\xi A) + Q(\xi^a\nabla_a V)\right]\psi \\
&= (\tfrac{\hbar}{i})^3\left[[\xi^c g^{ab}](2\nabla_{[a}\nabla_{c]}\nabla_b\psi) + [2g^{ab}\nabla_a\xi^c]\nabla_b\nabla_c\psi\right. \\
&\qquad \left. +[g^{ab}\nabla_a\nabla_b\xi^c]\nabla_c\psi + [g^{ab}\nabla_{(a}(\nabla_c\xi^c)]\nabla_{b)}\psi + [g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi\right] \\
&\quad -(\tfrac{\hbar}{i})\left[P(\pounds_\xi A) + Q(\xi^a\nabla_a V)\right]\psi \\
&= (\tfrac{\hbar}{i})^3\left[[\xi^c g^{ab}](R_{acb}{}^d\nabla_d\psi) + [2g^{ab}\nabla_a\xi^c]\nabla_{(b}\nabla_{c)}\psi\right. \\
&\qquad \left. +[g^{ab}\nabla_a\nabla_b\xi^c]\nabla_c\psi + [g^{ab}\nabla_{(a}(\nabla_c\xi^c)]\nabla_{b)}\psi + [g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi\right] \\
&\quad -(\tfrac{\hbar}{i})\left[P(\pounds_\xi A) + Q(\xi^a\nabla_a V)\right]\psi
\end{aligned}
$$

where we have introduced the Riemann tensor $R_{abc}{}^d$ for convenience

$$
\begin{aligned}
&= (\tfrac{\hbar}{i})^3\left[[2g^{a(b}\nabla_a\xi^{c)}]\nabla_b\nabla_c\psi + [\xi^c g^{ab}R_{acb}{}^d + g^{ab}\nabla_a\nabla_b\xi^d]\nabla_d\psi\right. \\
&\qquad \left. +[g^{ab}\nabla_{(a}(\nabla_c\xi^c)]\nabla_{b)}\psi + [g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi\right] \\
&\quad -(\tfrac{\hbar}{i})\left[P(\pounds_\xi A) + Q(\xi^a\nabla_a V)\right]\psi \\
&= (\tfrac{\hbar}{i})^3\left[[2g^{a(b}\nabla_a\xi^{c)}]\nabla_b\nabla_c\psi\right. \\
&\qquad +[g^{ab}(R_{acb}{}^d\xi^c + \nabla_a\nabla_b\xi^d)]\nabla_d\psi + [g^{ab}\nabla_{(a}(\nabla_c\xi^c)]\nabla_{b)}\psi \\
&\qquad \left. +[g^{ab}\nabla_a\nabla_b(\nabla_c\xi^c)]\psi\right] \\
&\quad -(\tfrac{\hbar}{i})\left[P(\pounds_\xi A) + Q(\xi^a\nabla_a V)\right]\psi
\end{aligned}
$$

My n^{th}-generation copy of this page had unreadable sections. So, this page is my attempt to reconstruct the content. I hope that I have maintained the style of the presentation and that I did not introduce any errors. (Please check!)

This Heisenberg equation of motion is normally only written when ξ^a is a symmetry of the geometry of configuration space, *i.e.*, when $\pounds_\xi g^{ab} = \xi^m \nabla_m g^{ab} - g^{mb}\nabla_m \xi^a - g^{am}\nabla_m \xi^b = 0 - 2g^{m(a}\nabla_m \xi^{b)} = 0$.

[In other words, ξ^a is a *Killing vector* since

$$0 = 2g^{m(a}\nabla_m \xi^{b)} = 2\nabla^{(a}\xi^{b)} \qquad \text{(Killing's equation)}.$$

Furthermore, if ξ^c is a Killing vector, then $\nabla_c \xi^c (= g_{ab}\nabla^{(a}\xi^{b)})$ vanishes.

Lastly, if ξ^a is a Killing vector, then $R_{acb}{}^d \xi^c + \nabla_a \nabla_b \xi^d = 0$. To see this, begin with the definition of the Riemann tensor

$$R_{abc}{}^d \eta_d = \nabla_a \nabla_b \eta_c - \nabla_b \nabla_a \eta_c.$$

If η_c is a Killing vector, then, using Killing's equation $\nabla_{(a}\eta_{c)} = 0$, we have

$$R_{abc}{}^d \eta_d = \nabla_a \nabla_b \eta_c + \nabla_b \nabla_c \eta_a.$$

By index-substitution,

$$R_{bca}{}^d \eta_d = \nabla_b \nabla_c \eta_a + \nabla_c \nabla_a \eta_b,$$

$$R_{cab}{}^d \eta_d = \nabla_c \nabla_a \eta_b + \nabla_a \nabla_b \eta_c.$$

Forming the sum "abc"+"bca"−"cab", we obtain

$$(R_{abc}{}^d + R_{bca}{}^d - R_{cab}{}^d)\eta_d = 2\nabla_b \nabla_c \eta_a.$$

Since $R_{[abc]}{}^d = 0$ and $R_{abc}{}^d = -R_{bac}{}^d$ implies that $R_{abc}{}^d + R_{bca}{}^d + R_{cab}{}^d = 0$, we find that

$$(-2R_{cab}{}^d)\eta_d = 2\nabla_b \nabla_c \eta_a.$$

Hence,

$$(-R_{cab}{}^d)\eta_d = \nabla_b \nabla_c \eta_a.$$

Playing with positions of the contracted-indices on the left-hand side,

$$-R_{cabd}\eta^d = \nabla_b \nabla_c \eta_a.$$

Since $R_{abcd} = R_{cdab}$,

$$-R_{bdca}\eta^d = \nabla_b \nabla_c \eta_a.$$

Raising the a-index,

$$-R_{bdc}{}^a \eta^d = \nabla_b \nabla_c \eta^a.$$

Hence, for a Killing vector ξ^a,

$$R_{acb}{}^d \xi^c + \nabla_a \nabla_b \xi^d = 0. \qquad]$$

In this case, the highest order term [*i.e.* the $(\frac{\hbar}{i})^3$ term] on the right above vanishes, and the formula reduces to

$$(HP(\xi) - P(\xi)H)\psi \quad = \quad -(\tfrac{\hbar}{i})\left[P(\pounds_\xi A) + Q(\xi^a \nabla_a V)\right]\psi$$

which, of course, reduces to the familiar formula in standard examples (*e.g.*, particle in Euclidean space). These calculations are merely intended to illustrate, for special cases, the general formula (63).

21. The Role of Observables. (Classical Mechanics)

Our approach heretofore has been to introduce the states before the observables. Thus, in classical mechanics, observables were functions on the space of states; in quantum mechanics, operators on the space of states—in both cases, observables were defined in terms of their relationship with states. It is natural to proceed in this way, for one thinks of a system as possessing a state (independently of any observations which might be made on it), with the observations representing things which are applied to these states to give us information about the state. One could, however, imagine an alternative point of view. One could regard states, not as attributes of a system somehow handed down from above, but rather as something which must be determined by means of observations. In other words, one could take a more observer-oriented point of view, in which the things which involve us observers directly—the observables—play the fundamental role. This point of view would be reflected within the mathematical formulation by a formalism in which the observables are introduced as the fundamental objects, and the states only as subsidiary objects defined in terms of the observables. As an example, we first carry out such a reformulation for classical mechanics.

We suppose that we are given some system to study. We have some mechanism (*e.g.*, hitting with a stick) by which we can cause our system to change state, so we can experiment with various states. We are to be furnished, furthermore, with an enormous basket full of observables, where we think of an observable as a box from which there protrudes a probe (which can be brought in contact with the system), and which has a meter (which reads some value when the system is in interaction with the observable). Our job is to introduce a description of what is going on in terms of the meter readings of the observables.

One first, by trial and error, makes the following basic observation: if, to the system, observable A, then B, and then A are applied in quick succession (so the system does not evolve appreciably between observables), then the readings obtained for the two A observations coincide. We interpret this to mean that the application of our observables does not disturb the state of the system (*i.e.*, the application of B leaves invariant the A-reading, for all As, and hence leaves the state alone). It follows that the application of several observables in quick succession is independent of the order in which the observables are applied.

Two observables, A and A', will be said to be *equivalent* if they always give the same result, applied to our system (more precisely, if the application of A and A' in quick succession yields the same reading on the two meters). We write equivalence of observables with an equals sign.

We lose nothing in generality, yet simplify the discussion, by assuming that our basket of observables has three properties. A *constant observable* is one whose meter reading is always the same number (no matter what the system is doing). We assume that, for each real number, there is in our collection of observables a constant observable which always yields that number. (If some constant observable were missing from our collection, it would be easy to construct one to add to the collection. Just take an empty box, and paint on it a dial and needle always reading, for example, "3".) Now let A and B be two observables. We define a new observable, obtained by applying A and B to the system in quick succession, and adding the readings on the two meters. This new observable, called the *sum of observables* A and B, will be written $A + B$. Similarly, the *product of two observables*, AB, is another observable. We now assume that our collection of observables contains, along with any two observables, both their sum and their product. (If some sum, *e.g.*, $A+B$, were missing from the collection, we could always construct, from A and B, $A + B$ and add it to our collection). Note that addition and multiplication of observables sat-

isfy the same commutative, associative, and distributive laws that addition and multiplication of real numbers do.

An **algebra** $\mathcal{A}$ is a vector space on which one or more bilinear products is defined. (Thus, let $\mathcal{A}$ be a vector space on which products AB, $A \times B$, and $[A, B]$ are defined, satisfying $(A+cB)C = AC+cBC$, $A(B+cC) = AB+cAC$, for c a number, and A, B, C in $\mathcal{A}$, and similarly for the other two products. Then $\mathcal{A}$ is an algebra.)

The discussion above can now be summarized by the statement that we might as well assume that our collection $\mathcal{A}$ of observables forms an algebra (with just one product, which is also associative and commutative).

A *state* is to be characterized in terms of what the observables have to say about it. That is to say, a **state** is a mapping $\sigma : \mathcal{A} \to \mathbb{R}$, from our algebra of observables to the reals, satisfying the following conditions:

1. If C is the constant observable with value c, then $\sigma(C) = c$.
2. If A and B are observables, then $\sigma(A + B) = \sigma(A) + \sigma(B)$.
3. If A and B are observables, then $\sigma(AB) = \sigma(A)\sigma(B)$.

These three conditions, of course, reflect the operational meaning of *constant observables, addition of observables, and multiplication of observables.*

This is nearly the end of the story. From the algebra of observables, one introduces the states. We have still to ask, however, whether or not we obtain, in this way, the "correct" collection of states. The answer depends on whether or not we have enough observables in our original collection $\mathcal{A}$. For example, if $\mathcal{A}$ consisted of just the constant observables, then we would obtain, by the construction above, just one state. Suppose, however, that all observables were included in our collection. Since we "really know" that our system has a phase space, and the observables are scalar fields on phase space, this amounts to the assumption that for every scalar field on phase space there is an instrument which measures that scalar field. We can then ask: *does our construction above yield the correct collection of states?* First note that every point of phase space does indeed define a mapping from observables to the reals (namely, evaluation of the scalar field at the point). Hence, every point of phase space does define a state according to our construction. What needs to be shown is that no additional states—no points of phase space—have been introduced. This is a consequence of the following:

Theorem. Let M be a manifold, and let $\mathcal{A}$ be the algebra of scalar fields on M. Let $\sigma : \mathcal{A} \to \mathbb{R}$ satisfy the three conditions above for a state. Then there exists a point p of M such that $\sigma(A) = A(p)$ for every A in $\mathcal{A}$.

"proof": the only proof I know requires the Whitney embedding theorem, a standard but rather technical result in differential geometry. Since a proof in the special case $M = \mathbb{R}^n$ illustrates the essential idea, and since a proof in this special case, together with the embedding theorem, yields easily a proof in general, we shall assume $M = \mathbb{R}^n$.

Let σ be a mapping as above, and consider the scalar fields $x^1, \ldots, x^n$ on M. Set $\underline{x^1} = \sigma(x^1), \ldots, \underline{x^n} = \sigma(x^n)$. Let p the point with coordinates $(\underline{x^1}, \ldots, \underline{x^n})$. Then, for f any scalar field on M, we can write $f = \underline{f} + f_1(x^1 - \underline{x^1}) + \cdots + f_n(x^n - \underline{x^n})$, where $\underline{f}$ is a constant scalar field, and $f^1, \ldots, f^n$ are scalar fields. Then $\sigma(f) = \sigma(\underline{f}) + \sigma(f_1)(\sigma(x^1) - \sigma(\underline{x^1})) + \cdots + \sigma(f_n)(\sigma(x^n) - \sigma(\underline{x^n})) = \sigma(\underline{f})$, where, in the last step, we have used that $\sigma(x^i) = \underline{x^i} = \sigma(\underline{x^i})$. Thus, $\sigma(f) = \sigma(\underline{f}) = \underline{f} = f(p)$, completing the proof.

Thus, if we have all the observables in our original collection, we can construct the space of states as above, and indeed obtain precisely the correct space of states. A similar conclusion holds even if original algebra includes fewer observables. For example, it easy to check, using the same methods above, that,

if $\mathcal{A}$ is the algebra of algebraic observables, then we obtain the correct space of states. Note, furthermore, that, if $\mathcal{A}$ includes all the configuration and momentum observables, then, since $\mathcal{A}$ is an algebra, $\mathcal{A}$ necessarily includes all algebraic observables. Then, so long as $\mathcal{A}$ includes all configuration and momentum observables, precisely the correct space of states emerges from the construction above.

22. The Role of Observables. (Quantum Mechanics)

Ideally, one would now like to repeat, for quantum mechanics, what we did in the previous section for classical mechanics. That is to say, one would like to introduce a collection of observables (measuring instruments) on a quantum system, subject to some physically reasonably conditions regarding the action of the instruments on the system, and then define the states in terms of the observables. In this way, one might expect both to obtain a more operational setting for quantum mechanics, and to provide more direct physical interpretations for certain objects which appear in the mathematical formalism of quantum mechanics. It appears, however, that certain difficulties intervene to prevent a simple and direct treatment of quantum mechanics along these lines. In this section, we discuss these difficulties.

We must first consider the mechanism by which observables act on states. Within the formalism of quantum mechanics, an observable is a self-adjoint operator on the space of states. What we must do is relate the mathematical action of this operator to the physical action of the observing instrument on the state. There are, apparently, two possible such relationships.

1. *Active observables.* One wishes to apply to the state the observable associated with the self-adjoint operator A. The meter on the instrument reads an eigenvalue of A, *i.e.*, a number such that $A\underline{\psi} = a\underline{\psi}$ for some nonzero state $\underline{\psi}$. During the observation, the system suddenly changes its state so that, after an observation yielding the value a, the system is in the corresponding eigenstate (*i.e.*, the $\underline{\psi}$ above). Even though the initial state ψ, as well as the self-adjoint operator A, are known, the final state into which the system passes is not uniquely determined. Rather, what is determined is the probabilities for the various eigenvalue-eigenstate pairs.

2. *Passive observables.* The application, to the state ψ, of the observable associated with the self-adjoint operator A yields (as the reading of the instrument) the expectation value of A in our state, *i.e.*, the real number $\langle\, \psi \,|\, A\psi \,\rangle$. The effect of this measurement on the state of the system can be made arbitrarily small.

I believe that it is true to say that, given A, one can construct measuring instruments, of either the active or the passive type, corresponding to A. The measuring instruments we introduced in classical mechanics were passive, although one could, of course, have introduced also active instruments (which would change the state of the system). Of course, one could, if it should turn out to be convenient, simply require that the instrument he uses be of one or the other type. In particular, one often treats quantum mechanics using only active instruments. The possibilities for defining states in terms of observables depend on which type of observing instruments one chooses to use. We shall consider the results of using both types.

The disadvantage of using active instruments is that they interfere, in a significant and uncontrollable way, with the state being observed. In fact, the original state is for all practical purposes destroyed by the making of a single active observation. However, a vestige of the original does survive in that the original state determines the probability distribution for the various possible eigenvalues which could emerge from the observation. Unfortunately, a single observation on a single system yields, not a probability distribution, but a single eigenvalue. It is clearly impossible to describe a state in terms of observations on that state in the presence of this strong interference of the observing instrument with state itself.

There is, however, a method for avoiding at least certain of these difficulties. Let us imagine that we have been given a method for preparing the system so

as to be in a given state. We then make up an infinite collection of copies of our original of copies of our original system, all in the same state. This collection is called an **ensemble**. One now imagines applying observing instruments, not to a single system (in the single state), but to an ensemble. This is done by making the observation on some infinite sub-collection of the systems forming the ensemble, leaving, at the same time, an infinite collection of systems in the ensemble unobserved. (For example, one could observe every other system in the ensemble.) For each system observed, one records the number (eigenvalue) determined by the observing instrument. The observed systems can then be discarded (their states having been ruined). There remains, however, an infinite collection of systems which have not been disturbed, *i.e.*, an ensemble for further study. Thus, by introducing ensembles rather than states, one avoids the difficulties associated with the interference of observations with the state of the system. One must pay for this device, however, in that one must, in a sense, already know what the states are, for, to prepare an ensemble, one must have in hand a repeatable prescription for putting the system in a given state.

In any case, the information we obtain by applying an observable to an ensemble is a **probability measure** on the real line (*i.e.*, a positive measure such that the total measure of the line is one).

Thus, one would attempt to proceed as follows. One would introduce an algebra of (active) observables. States would be defined as mappings from this algebra to probability measures on the real line, such that the mapping satisfy certain conditions. One would then have to show that the correct space of states is obtained in this way. Unfortunately, it appears to be very difficult to make such a program work, in the sense that one finds suitable structure on the observables and suitable additional conditions for the definition of a state, such that the structure and conditions can be motivated physically, and such that the resulting collection of states agrees with the space of states of quantum mechanics.

It is perhaps not too surprising that the use of active observables should lead to difficulties. The program of defining states in terms of observables works extremely well in classical mechanics, and in that case the observables are in every sense passive. One would naturally expect that the passive observables would be the most natural to use also in quantum mechanics.

Let us attempt, then, to describe states in terms of expectation values for self-adjoint operators, *i.e.*, in terms of passive observables. It is perhaps not unreasonable to ask that included in our collection of observables be all configuration and momentum observables. We begin with a mathematical question: *does knowledge of the numbers* $\langle\, \psi \,|\, Q(\alpha)\psi \,\rangle$ *and* $\langle\, \psi \,|\, P(\xi)\psi \,\rangle$ *for all* α *and* ξ^a *determine* ψ*?* If not, our goal is hopeless; if so, we can ask how this determination comes about. Set $\psi = \phi e^{i\lambda}$, where ϕ and λ are real scalar fields. (For simplicity, we take ψ as a scalar field.) Then

$$\langle\, \psi \,|\, Q(\alpha)\psi \,\rangle = \int \overline{(\phi e^{i\lambda})}\alpha\phi e^{i\lambda} = \int \phi^2\alpha \tag{64}$$

$$\begin{aligned}
\langle\, \psi \,|\, P(\xi)\psi \,\rangle &= \int \overline{(\phi e^{i\lambda})}(\tfrac{\hbar}{i})(\xi^a\nabla_a + \tfrac{1}{2}\nabla_a\xi^a)\phi e^{i\lambda} \\
&= \int (\tfrac{\hbar}{i})\phi e^{-i\lambda}(e^{i\lambda}\xi^a\nabla_a\phi + i\phi e^{i\lambda}\xi^a\nabla_a\lambda + \tfrac{1}{2}\phi e^{i\lambda}\nabla_a\xi^a) \\
&= \int (\tfrac{\hbar}{i})\phi\xi^a\nabla_a\phi + \int \hbar\phi^2\xi^a\nabla_a\lambda + \int \tfrac{1}{2}(\tfrac{\hbar}{i})\phi^2\nabla_a\xi^a \\
&= \int \tfrac{1}{2}(\tfrac{\hbar}{i})(2\phi\xi^a\nabla_a\phi + \phi^2\nabla_a\xi^a) + \int \hbar\phi^2\xi^a\nabla_a\lambda \\
&= \int \tfrac{1}{2}(\tfrac{\hbar}{i})\nabla_a(\phi^2\xi^a) + \int \hbar\phi^2\xi^a\nabla_a\lambda
\end{aligned}$$

$$= \int \hbar\phi^2\xi^a\nabla_a\lambda$$
$$= \hbar\langle\,\psi\,|\,Q(\xi^a\nabla_a\lambda)\psi\,\rangle \tag{65}$$

Knowledge of the expectation value of $Q(\alpha)$ for all α determines, by (64), ϕ. Then the expectation value of $P(\xi)$ for all ξ^a determines, by (65), $\nabla_a\lambda$, and hence λ up to a constant. Thus, $\psi = \phi e^{i\lambda}$ is determined up to an overall phase (which is all the wave function is ever determined up to, anyway). Our program should, in principle, succeed.

Denote by $\mathcal{Q}$ the collection of all (passive) configuration observables, and by $\mathcal{P}$ the collection of all (passive) momentum observables. One knows, operationally, how to multiply an element of $\mathcal{Q}$ or an element of $\mathcal{P}$ by a real number, and how to add two elements of $\mathcal{Q}$ or two elements of $\mathcal{P}$. Thus, each of $\mathcal{Q}$, $\mathcal{P}$ has the structure of a vector space. A *state* is to consist of linear mappings $\sigma : \mathcal{Q} \to \mathbb{R}$ and $\sigma : \mathcal{P} \to \mathbb{R}$. We wish to impose additional conditions on these mappings in order that $\sigma(\alpha) = \langle\,\psi\,|\,Q(\alpha)\psi\,\rangle$ and $\sigma(\xi) = \langle\,\psi\,|\,P(\xi)\psi\,\rangle$ for some state ψ. It follows from the linearity of σ that there exist scalar and vector fields μ and μ_a on configuration space such that

$$\sigma(\alpha) = \int \alpha\mu \tag{66}$$

$$\sigma(\xi) = \int \xi^a\mu_a \tag{67}$$

for all α and ξ^a. (We should, more precisely, allow μ and μ_a to be *distributions* on configuration space. Thus, for example, a scalar distribution is normally defined as a linear mapping from scalar fields on a manifold to reals. In order to avoid technical complications, however, we shall restrict consideration to the case when μ and μ_a are ordinary tensor fields.) We wish to define ϕ by $\phi^2 = \mu$, and for this we need $\mu \geq 0$. We can state this condition in terms of our operators as follows: if $\alpha \geq 0$, then $\sigma(\alpha) \geq 0$. We wish to define λ by $\hbar\phi^2\nabla_a\lambda = \mu_a$, which requires that $\phi^{-2}\mu_a$ be a gradient. This condition is equivalent to the requirement that, whenever ξ^a is such that $\nabla_a(\phi^2\xi^a) = 0$, $\sigma(\xi) = 0$, *i.e.*, to the requirement that whenever ξ^a is such that $\sigma(\xi^a\nabla_a\alpha) = 0$ for all α, $\sigma(\xi) = 0$. With these two conditions, we can recover ϕ and λ, and hence the state $\psi = \phi e^{i\lambda}$, from the mappings σ.

To summarize, we introduce two vector spaces, $\mathcal{Q}$ and $\mathcal{P}$, of observables. A state is then defined as a pair of linear mappings, $\sigma : \mathcal{Q} \to \mathbb{R}$ and $\sigma : \mathcal{P} \to \mathbb{R}$, satisfying two conditions:

1. If $\alpha \geq 0$, then $\sigma(\alpha) \geq 0$.
2. If $\sigma(\xi^a\nabla_a\alpha) = 0$ for all α, then $\sigma(\xi) = 0$.

Then, roughly speaking, for every such state there is a unique wave function ψ on configuration space such that $\sigma(\alpha) = \int \bar{\psi}\alpha\psi$ and $\sigma(\xi) = \int \bar{\psi}(\frac{\hbar}{i})(\xi^a\nabla_a + \frac{1}{2}\nabla_a\xi^a)\psi$.

Note that our two conditions in the definition of a state are indeed satisfied in the classical case. (Classically, the second condition amounts essentially to the statement that vector fields are defined as derivations on scalar fields, so the zero derivation defines the zero vector field). Why were we able to avoid explicit mention of such conditions in the classical case? The reason is that we were able there to introduce a far stronger condition in the definition of states, namely, the product condition: $\sigma(AB) = \sigma(A)\sigma(B)$. One might imagine, therefore, that matters could be simplified in the quantum case by the introduction of a similar product condition. Of course, it is easy, operationally, to introduce the product of two (passive, quantum) observables. Given instruments which measure A and B, one constructs an instrument which measures first A and then B, and multiplies. (Since passive observables do not affect the states, this

construction is independent of order.) The difficulty, however, is this: if A and B are self-adjoint operators, then, in general,

$$\langle\, \psi \,|\, A\psi \,\rangle \langle\, \psi \,|\, B\psi \,\rangle \neq \psi AB\psi \tag{68}$$

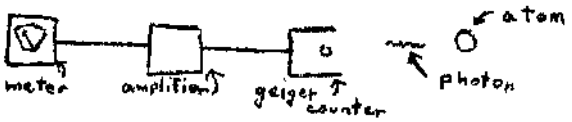
The left hand side is the number associated with the "product observable" constructed above. The right side is the expectation value associated with the product of operators. In short, the operational procedure of taking products of observables does not correspond to the mathematical procedure of taking products of operators. The incorporation of operational products into our quantum observables thus adds essentially nothing new.

Thus, we indeed obtain a formulation of quantum mechanics in which states are described in terms of (passive) observables. Although the two conditions we require in the definition of a state (that $\sigma(\alpha) \geq 0$ if $\alpha \geq 0$, and that $\sigma(\xi) = 0$ if $\sigma(\xi^a \nabla_a \alpha) = 0$ for all α) are perhaps not terribly unnatural physically, the second certainly seems rather artificial. One could easily invent other conditions of the same general type which are not imposed. There is, however, another—perhaps even more unpleasant—feature of this formulation. In the classical case, we could think of our observables as measuring instruments, instruments which we knew how to apply to our system, but about which no further information was required. This is not at all the case for our quantum observables. Firstly, we need to know which instruments measure configuration and which momentum observables. That is to say, each instrument must have, inscribed on its side, either "$\mathcal{Q}$" or "$\mathcal{P}$". In fact, we need, for the construction of states above, much more than this. We also need to know, for each instrument which measures a configuration observable, which scalar field α on configuration space this instrument measures (the expectation value of), and, for each instrument which measures a momentum observable, which vector field ξ^a is observed. This additional information is necessary in order that the construction of states can be carried out. On the other hand, no operational procedure is given for obtaining this additional information, *e.g.*, no procedure for discovering what an instrument measures by taking it apart. In fact, the introduction of such observables requires already a knowledge of what configuration space is—but configuration space determines already the quantum states.

To summarize, because the measurement process in quantum mechanics is apparently more subtle than that in classical mechanics, the simple formulation of classical mechanics in terms of observables apparently does not carry over in a natural and equally simple way into quantum mechanics.

23. The Interpretations of Quantum Mechanics

It is conventional to attempt to formulate what are called *interpretations of quantum mechanics*. It is natural to ask, firstly, why such an interpretation is needed. Apparently, it is because the objects which appear in the mathematical formulation of quantum mechanics (wave functions, self-adjoint operators, *etc.*) are related in a rather tenuous way to the things (*e.g.*, dial readings) which human observers actually see. Consider an example. An atom in some excited state is to decay, emitting a photon. The photon is to ionize an atom in a geiger counter, so the resulting electron will, because of an applied electric field, set up a cascade of electrons in the counter. The resulting electric signal will, after being sent through an amplifier, activate the needle of a meter. We can begin this system with the original atom in an excited state, and ask how the system will evolve. This is a question with which we can, at least in principle, deal, within the context of quantum mechanics. We are to introduce the configuration space of this entire system (consisting of the atom, the geiger counter, the amplifier, and the meter). On this configuration space, we are to introduce an appropriate differential operator, the Hamiltonian. We are then to introduce a certain wave function ψ describing the initial state of the system. Schrodinger's equation will then determine ψ at later times.



Clearly, the program above is so formidable that one is not likely to carry it out in practice. One can, however, gain insight into what a solution would look like by considering a simpler problem having roughly similar features. Consider a system with 2-dimensional configuration space, the xy-plane. Let the classical Hamiltonian be

$$H = {p_x}^2 + x^2 + {p_y}^2 + y^2 + f(x - y)$$

where f is some function of one variable. Thus, when f is zero, we have two independent one-dimensional harmonic oscillators, f represents a coupling between two such oscillators. We write down the quantum description of this system, and begin in the state $\psi(x, y) = \psi_g(x)\psi_e(y)$, where $\psi_g(x)$ is the ground state, and $\psi_e(y)$ some excited state, of a one-dimensional harmonic oscillator. We now allow this state to evolve, by Schrodinger's equation, and see what happens. What happens is exactly what one expects to happen: the energy "leaks" slowly, because of the coupling, from one oscillator to the other. That is to say, after a short time, the first oscillator is nearly in its ground state, but with a small admixture of excited states. As time goes on, the amount of this admixture slowly increases.

We can think of this simple system as analogous to the more complicated system which preceded it. Initially, the system was described by a wave function such that the atom was in an excited state, and the needle on the meter reads zero. As time goes on, the component of the wave function of the atom corresponding to the ground state increases, while the component of the wave function of the needle corresponding to a nonzero meter reading also increases. That is to say, one would expect that the state of our system will evolve, slowly

and continuously, from one in which the atom is excited and the meter reads zero, to one in which the atom has decayed and the meter reads nonzero.

On the other hand, one can actually buy, *e.g.*, from Sears, a geiger counter, amplifier, meter, and excited atom. One sets up the experiment, and watches the needle as the system evolves. What one actually sees, of course, is that the needle remains at zero for a while, then jumps to some value. One often says that, at the instant the needle jumped, the atom decayed, and the photon was received by the geiger counter, amplified, and recorded by the meter.

We now have two descriptions of this experiment. In one, via the formalism of quantum mechanics, the wave function of the needle continuously evolves from that corresponding to a zero reading to that corresponding to a finite reading. In the other, via direct observation, the needle is seen, not in some "smeared out state", half reading zero and half something else, but is rather seen to remain at zero for a while then jump suddenly. The problem of interpretation, as it bears on this example, is to somehow bring together these two descriptions. That is to say, the problem of interpretation is to make more sharp the relationship between the objects which appear in the formalism of quantum mechanics and the things observers actually see.

It is not at all uncommon in physics (perhaps it is even characteristic!) that the objects at the center of the mathematical description are not the first and most obvious things observers see. Special relativity is a good example. Fundamental in the mathematical description of special relativity are the events (instantaneous occurrences at a point), which are assembled into the four-dimensional space-time manifold. Yet, although each of us lives (ignoring gravity) in the four-dimensional space-time manifold of special relativity, we do not experience any four-dimensional manifold stretching out before us. Instead, we experience three-dimensional space, and the passage of time. Why, then, does one not feel a need to acquire an "interpretation" of special relativity?

It is my impression that an interpretation is as necessary in special relativity as in quantum mechanics, but that, in the case of the former, an interpretation is available and widely accepted, so that its role as an interpretation normally goes unemphasized. This interpretation is that each of us (our past, present, and future) is represented by a world-line in space-time, that what we experience is light rays reaching our world-line from distant events, objects whose world-lines meet ours, *etc.* Within this framework—a framework in which the observer is incorporated directly—one can, in some sense, account for what we as observers actually see.

It is perhaps not the whole truth to say that those experiences for which special relativity might give account are in fact accounted for by special relativity. It is, for example, not so easy to account, within special relativity, for the fact that one cannot go back and actively experience his past. However, one perhaps takes the point of view that these finer details are somehow to be described in terms of the detailed physics of what goes on in the brain. One might also say, based on this discussion of special relativity, that the purpose of an interpretation is, not really to explain anything, but rather to put one at ease while he works within the mathematical formalism.

We return to quantum mechanics. As discussed above, there is a sense in which quantum mechanics is inadequate. On the other hand, there is certainly a sense in which quantum mechanics is far more than adequate: *it makes beautiful and accurate predictions in remarkable agreement with Nature.* One would therefore not like to change in a substantive way the structure of quantum mechanics, *e.g.*, change its mathematical formalism. What one would like to do is to look at quantum mechanics from a different point of view, or at least utter some reassuring words about it, so that the apparent inadequacy seems less serious. That is to say, what one would like to do is formulate an interpretation of quantum mechanics.

24. The Copenhagen Interpretation

The most widely accepted interpretation of quantum mechanics is what is called the **Copenhagen interpretation**.

The idea of the Copenhagen interpretation, as I understand it, is that quantum mechanics is fine as far as it goes, but that it does not go far enough. What is missing is the relationship between the wave function and what observers see. The Copenhagen interpretation attempts to provide this additional link.

One first notices that, if you ask an observer what it was that he actually saw, he will say "I saw the meter read 3.5.", or "I saw the light go on.", but never "The wave function of the needle on the meter was the following superposition...". Thus, this additional link is to somehow tie together quantum phenomena (*i.e.*, wave functions) and classical phenomena (*i.e.*, points of configuration space).

Consider again the experiment on page 63. We wish to regard the meter classically, and the original atom quantum mechanically. We therefore introduce a break at some point between the chain of instruments linking the atom to the meter, *e.g.*, between the atom and the geiger counter. Everything up to the break (in the case, just the atom itself) will be treated quantum mechanically, and everything beyond the break (in this case, the geiger counter, amplifier, and meter) will be treated classically. Put more formally, we introduce a classical phase space to describe the states of the meter-amplifier-geiger counter system, and a space of quantum states (*i.e.*, complex functions on an appropriate configuration space) to describe the original atom.

The next step is to introduce a coupling of these two systems, across the break. This is done by assigning a self-adjoint operator A to the quantum system, and, at the same time, assigning a state of the classical system to each eigenvalue of A. The prescription for making these assignments is not specified by the interpretation, but the assignments are to be so made that, in the end, one obtains an appropriate description of this measurement process. The essential assumption is that one can introduce a break in this chain of instruments, and describe the information flow across the break in terms of a self-adjoint operator A on the quantum system, and states of the classical system labeled by eigenvalues of A.

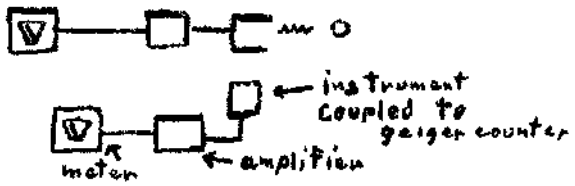
The break comes into play during what is called *process of measurement*. The following then takes place. The Schrodinger equation for the quantum system, and the Hamilton equation for the classical system are, for a moment, suspended. During this suspension the quantum system is thrown into an eigenstate of A, and, simultaneously, the classical system is thrown into the state associated with the corresponding eigenvalue of A. The process of measurement is thus a process of sudden change in state. In this way, the formalism of quantum mechanics is to be brought into contact with the (classical) things people, as observers, see.

The particular eigenstate of A into which the quantum state is thrown during the process of measurement is not specified precisely. Rather, what can be determined is the probability distribution for various of the probabilities. Thus, there is, in a sense, a loss of determinism in the Copenhagen interpretation, associated with the process of measurement. It is essentially at this point that probability is brought into quantum mechanics.

One might object to the interpretation above on, among others, the following grounds. No prescription has been given for where to introduce the break between the quantum and classical systems. Thus, in our example, one could as well have drawn the break between the geiger counter and amplifier, or between the amplifier and meter. It is, so I understand, part of the lore of the Copenhagen interpretation that the result (*i.e.*, the resulting probability distribution for the meter readings) is independent of this choice. It is not completely clear how one formulates a precise statement along these lines, although indications are (*e.g.*, from simple examples) that something of this general type is true.

Thus, one is apparently free to introduce the break at any point of the chain with the same result, *i.e.*, with the same probability distribution for (classical) meter readings.

One might now object on somewhat different grounds. Since one is free to introduce the break wherever one wishes, it is prescribed, within the interpretation, whether or not a given link in the chain of measurement instruments is quantum. Thus, for example, the geiger counter could be regarded either as a classical or as a quantum system, depending on where the (arbitrary) break was introduced. The Copenhagen answer would be this. If you are concerned about the geiger counter, make a measurement on it. Thus, one would introduce a tributary in the chain of measurements, with additional instruments looking in particular at the geiger counter.



Now one has a new system of linked instruments. One is supposed to apply to it the Copenhagen prescription, *i.e.*, one is to introduce a break in the chain, with the two sides linked by a self-adjoint operator, *etc.* The result will be a probability distribution for classical results (at the end of the chain), a distribution independent of where the break was drawn. In any case, one will obtain an unambiguous result. Whether one chooses to interpret this result as some statement about whether the geiger counter is quantum or classical is certainly one's option. But, in any case, no difficulty arises from the failure of the Copenhagen interpretation, in the first experiment, to specify whether the geiger counter was "classical" or "quantum".

The claim of the interpretation is just that

if you specify to it with sufficient precision what you will do in making a measurement, it will tell you the probability distribution for the results of the measurement.

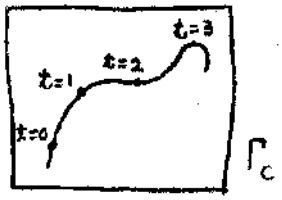
It does precisely this in every case. It does not have to deal with what "really happens". It is, in some sense, "consistent".

One could, equally well, have expressed this interpretation in a somewhat different way. One could as well have asserted, at some point, that the entire system is in fact subject to the laws of quantum mechanics. That is to say, one could have asserted that, in fact, there is no break in the system. Then the Copenhagen prescription could have been regarded merely as the rules for carrying out the calculation giving the probability distribution for the results of the measurement. That is to say, one could have presented matters so that the break business was presented as a computational tool rather than a statement about what "really happens". Clearly, nothing of any real substance (*e.g.*, numerical conclusions) depend on this mode of formulation. This "modified Copenhagen interpretation" is, however, perhaps more natural sounding, for it requires only that we in our calculations do strange things (*e.g.*, introduce a break), rather than that Nature do them.

25. The Everett Interpretation

There is an alternative interpretation of quantum mechanics, called the **Everett interpretation**, which we now describe.

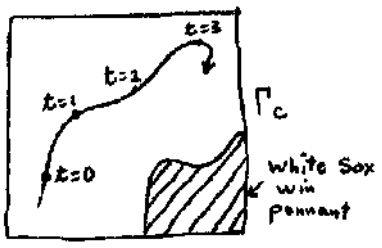
It is convenient to begin with some remarks about classical mechanics. The formalism of quantum mechanics was introduced originally, in order to describe rather implicitly defined things called *systems*. Let us, just to see what will result, attempt to push this formalism into a context broader than that for which it was originally intended. What we wish to do is to treat the entire Universe as a single system within the formulation of classical mechanics. Thus, we suppose that there is some configuration space C for our Universe. Let Γ_C be the corresponding cotangent bundle. We suppose that there is some scalar field H on this Γ_C, the Hamiltonian. In the beginning (*i.e.*, at time $t = 0$), the Universe occupied some point of phase space. The future evolution of the Universe is then determined by Hamilton's equation, *i.e.*, by the condition that the dynamical trajectory describing the evolution of the Universe be an integral curve of the Hamiltonian vector field. We emphasize that within this description is to be included the *entire* Universe—you and I, all observers, all observing instruments, everything.



One might feel uncomfortable about the above for the following reason. In our original formulation of classical mechanics, one imagined an external observer who, by manipulation and examination of the system under study, was able to assign it a configuration space and Hamiltonian. But the present context excludes such external observers. Thus, for example, consider a point of phase space through which the dynamical trajectory of our universe does not pass. What does it mean to speak of this "possible state of the Universe" when the Universe never, in fact, occupies that state?

It seems to me that the extrapolation involved here is, essentially, no better and no worse than the extrapolations one always makes in physics. Suppose, for example, that an external observer wishes to study a classical system. He has some means to manipulate the system, in order to study its states. Nonetheless, the system will, even with external manipulation, follow some path in its phase space, a path which will not reach every point of phase space. Thus, even in this case, one is introducing fictitious states which the system will never in fact possess. As a second example, in special relativity one begins by considering all possible events (occurrences having extension in neither space nor time), which are then to be assembled into a four-dimensional space-time manifold. But certainly there are events in space-time which no observer has witnessed. In each case, the mathematical description involves an extrapolation well beyond what is directly observed. Apparently, it is just the what we do physics to introduce, at the beginning, a framework which, if examined too closely, is unjustified.

The description of all that happens in the Universe is now in terms of the phase space of the Universe, and its dynamical trajectory. Thus, "possible occurrence" is another way of speaking of a certain region of phase space Γ_C.



For example, "the White Sox win the pennant" defines a certain region of Γ_C, namely, those states of the Universe in which the White Sox win the pennant. The question of whether this actually occurs is the question of whether or not the dynamical trajectory of our Universe actually passes through this region. In short, we are reduced to characterizing certain regions of Γ_C and asking whether or not the dynamical trajectory of the Universe actually passes through such regions. In short, we are reduced to characterizing certain regions of Γ_C and asking whether or not the dynamical trajectory of the Universe actually passes through such regions. We emphasize that you and I (who usually think of ourselves as the external observers) are now part of the system, so our "observations" are also described, within the framework above. For example, "I make a position measurement on a free particle, and obtain $x = 2$, $y = -3$, $z = 7$" simply describes a certain region of Γ_C, through which the dynamical trajectory of the Universe may or may not pass.

One might now object that we have not, through the discussion above, acquired any additional predictive power, since we will not, presumably, succeed in finding the phase space and Hamiltonian for the Universe, much less succeed in finding our dynamical trajectory. The first point is that neither have we lost any predictive power. Our decision to regard the entire Universe as a classical system does not negate the classical mechanics which has already been done. The second point is that our goal is not to eventually discover the phase space and Hamiltonian of the Universe and predict. Rather, we are trying to push classical mechanics to an extreme to see what sort of picture emerges. The resulting picture is relatively straightforward and not particularly unpleasant. (Perhaps one might be tempted to add to this last sentence "and unjustified". One could have done the same in special relativity.)

We restate the viewpoint above in different words. Let us imagine, for a moment, an external observer O of the Universe. He has described this system by a phase space and Hamiltonian, and complacently watches it follow its dynamical trajectory in phase space. We construct a sub-system and make observations on it. O notes with satisfaction that the dynamical trajectory of the Universe indeed enters that region of phase space. We formulate some sort of objection about predictive power. O notes with satisfaction that the dynamical trajectory of the Universe enters that region of phase space. In fact, all O ever does is note with satisfaction that the dynamical trajectory enters various regions. We might as well dispense with O.

The Everett interpretation of quantum mechanics amounts essentially to the discussion above, applied to quantum mechanics rather than classical mechanics. Thus, we imagine a configuration space C for the Universe, and a differential operator H, the Hamiltonian, on C. At time $t = 0$, the Universe begins in a certain state, *i.e.*, one has a certain density ψ on C. The future evolution of ψ is then determined by Schrodinger's equation. Included within this system are all observers, their instruments, *etc.* This is the picture which emerges if one tries to apply quantum mechanics to the system consisting of the Universe as a whole. The idea of the interpretation is simply to take this picture seriously.

When referring to a system, one speaks of something happening in that system, one normally means that some classical possibility has been realized. Suppose we watch a quantum system consisting of a geiger counter and atom, such that the photon resulting from the decay of the atom would be registered by the geiger counter. Our description of this system would be by a wave function which evolves by Schrodinger's equation. The wave function may, as it evolves, become large or small for certain classical possibilities (*i.e.*, in certain regions of configuration space), but nothing ever "happens" in the classical sense. The scheme above requires that we regard the Universe as a quantum system. Thus, from this viewpoint, nothing "happens" in the Universe (in the sense that no classical possibilities are actually realized). In other words, since implicit in quantum mechanics is a wave function rather than the occurrence of classical possibilities, we are forced, if we are to regard the Universe as a quantum system, to deny the occurrence, in the Universe, of classical possibilities.

One might object, at this point, that this denial is in disagreement with our everyday observations. We, apparently, do see classical possibilities "occur". If we watch the meter attached to the geiger counter, we do see, at some point, the needle jump. The response to this objection, from the present point of view, is that this impression of ours is simply another manifestation of the wave function of the Universe. Suppose that we, as observers, were watching a quantum system consisting of some rabbits, who have geiger counters and atoms, and who do various experiments. The rabbits may indeed formulate some sort of quantum theory in which classical possibilities seem to actually occur. We, however, would describe this system quite differently. We would say that there is an evolving wave function on the configuration space of the system. Whatever impressions the rabbits get would simply be a manifestation of that wave function. Thus, from the present viewpoint, we regard our everyday impression that classical possibilities actually occur as being a manifestation of the wave function of the Universe.

One could now formulate a second objection, that our impression that classical possibilities actually occur is not being explained by this framework, in the sense that we are not shown in detail how this impression is represented by the wave function of the Universe. The point is that this is something we will not be able to explain without a great deal more information about how Humans are constructed. But this is perhaps not so unpleasant, for there are many everyday human impressions which are not adequately described by physics, presumably because of a lack of information about the workings of people.

For example, we all observe that we cannot go back and actively re-live our pasts. This observation presumably falls under special relativity. Yet it is difficult to construct some theorem in special relativity which reflects this observation. Our understanding of this everyday observation would presumably improve if we understood better the internal workings of people. Yet, one can get along quite well in special relativity without such an understanding. A similar point is to be made within the Everett interpretation. That people feel that classical possibilities are actually realized simply reflects their internal construction (that, after all, is what we would say about the rabbits). It would be nice to understand this matter better, but we have too little information, and no particular need, to do so at present.

In the Everett interpretation, no classical possibilities ever actually "occur". Classical possibilities are reflected only in the wave function is a function on the space of such possibilities (*i.e.*, on configuration space).

In fact, this link between the formalism and what people actually experience can be made slightly stronger. Consider a region of the configuration space of the Universe in which the wave function is small. We can regard this region as representing classical possibilities which "for all practical purposes, do not occur". Consider now the experiment pictured on page 63. Because of the Hamiltonian, in the region of configuration space corresponding to the atom's

having decayed and the needle on the meter reading zero, the wave function is small. Thus, this classical possibility "for all practical purposes, does not occur". Similarly, if we think of the region of the configuration space of the Universe corresponding to the White Sox winning the pennant, and my impression that they did not, then one could expect that the Hamiltonian of the Universe is such that, in this region, the wave function is small. In short, one eliminates "a classical possibility occurs" as the link between the formalism and human experience. What replaces it is "in this region of configuration space, the wave function is small". In other words, one must learn to make statements about the Universe, not in the form "The following classical possibility ... occurs", but rather in the form "In the following region of configuration space ... the wave function is small". With a little practice, one can express himself equally well in this way.

We consider an example. Suppose we have some ordinary quantum system which contains a meter which, classically, can read either "A" or "B". By, *e.g.*, the Copenhagen interpretation, we determine that "the probability of reading A is 25%, while that of B is 75%". We wish to express this idea within the Everett language. The Copenhagen statement is to mean that, if this experiment were repeated many times, A would occur 25% of the time. This would not do as an Everett statement, however, because it refers to classical possibilities actually occurring.

We can, however, formulate this as follows: "In the region of configuration space corresponding to 1,000,000 of these instruments side by side, with the number reading 'A' either less than 240,000 or greater than 260,000, the wave function is small." Indeed, one would expect this to be the case. Suppose we construct a classical system which flips 1,000,000 pairs of coins, and records the number of times that both coins read heads. We quantize it, and ask what the wave function looks like after evolution through this process. One would, of course, find that the wave function will be small in the region of configuration space corresponding to "both heads" occurring less than 240,000 or more than 260,000 times. In a similar way, one formulates other statements about the Universe in terms of regions of configuration space in which the wave function is asserted to be small. Note that, whereas probability enters the Copenhagen interpretation "externally", it here enters though the internal structure of quantum mechanics itself.

In short, the Everett interpretation asks that one take quantum mechanics, as is, very seriously, and learns to live with the resulting picture. One gives up the notion of certain classical possibilities being realized in favor of the introduction of certain regions of configuration space in which the wave function is small. One carries out the same calculations, and transmits the same information, but in slightly different language. One obtains precisely the same description of the Universe that would be obtained by some external observer O. This O, however, would do nothing except look on with satisfaction as the wave function of Universe evolves. We might as well dispense with him. One does not need a classical framework in which to anchor quantum mechanics: one can just let quantum mechanics drift on its own.

Finally, one might object: "All this seems awfully philosophical and rather pointless." Imagine yourself in the following situation. You wake up one morning to discover that people always talk to each other by saying "In the region of configuration space corresponding to ... the wave function is small." That's just the way they always talk. You put up with this very confusing situation for a few days, and finally can't stand it anymore. You ask a friend to come in to see you. You say to him: "I want to reformulate quantum mechanics in such a way that classical possibilities actually occur in the Universe. I want to introduce smaller quantum systems, and observables, and breaks in the chain of instruments, on one side of which classical possibilities are actually realized. I want to modify, along these lines, the interface between quantum mechanics

and what human beings actually observe. It is true that, in this program, I cannot provide details of the internal workings of people, but this feature is also common in other areas of physics." After a pause, your friend replies: "All this seems awfully philosophical and rather pointless."

Robert Geroch

Differential Geometry

1972 Lecture Notes

MINKOWSKI
Institute Press

Robert Geroch
Enrico Fermi Institute
University of Chicago

© Robert Geroch 2013
All rights reserved. Published 2013

ISBN: 978-1-927763-06-3 (softcover)
ISBN: 978-1-927763-07-0 (ebook)

Minkowski Institute Press
Montreal, Quebec, Canada
http://minkowskiinstitute.org/mip/

For information on all Minkowski Institute Press publications
visit our website at http://minkowskiinstitute.org/mip/books/

Contents

1 Manifolds 1

2 Smooth Functions: Smooth Mappings 4

3 Vectors at a point 7

4 Linear Algebra 10

5 Tensors at a point 13

6 Tensor Fields 14

7 Lie Derivatives 16

8 Symmetrization and Antisymmetrization 21

9 Exterior Derivatives 22

10 Derivative Operators 26

11 Concomitants 29

12 Curvature 32

13 Metrics 35

14 Curvature Defined by a Metric 38

15 Smooth Mappings: Action on Tensor Fields 41

16 Bundles 46

17 The Tensor Bundles 49

18 Smooth Mappings: Action on Tensor Bundles 51

19 Curves 53

20 Integral Curves 55

21 The Lie derivative: Geometrical Interpretation 59

22 Lie Groups 61

23 Groups of Motions 63

24 Dragging Along 65

25 Derivative Operators: Interpretation in the Tensor Bundles 67

26 Riemann Tensor: Geometrical Interpretation 69

27 Geodesics 70

28 Submanifolds **71**

29 Tangents and Normals to Submanifolds **73**

30 Metric Submanifolds **74**

31 Tensor Fields and Derivatives on Submanifolds **77**

32 Extrinsic Curvature **80**

33 The Gauss-Codazzi Equations **83**

34 Metric Submanifolds of Co-Dimension One **85**

35 Orientation **87**

36 Integrals **89**

37 Stokes' Theorem **91**

38 Integrals: The Metric Case **94**

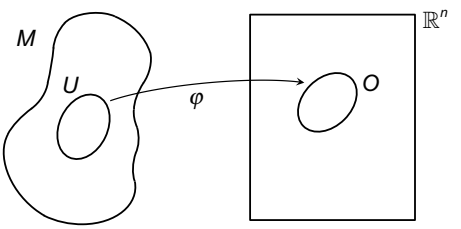
Appendix **97**

1. Manifolds

The arena on which all the action in differential geometry takes place is a mathematical object called a manifold. In this section, we shall define manifolds and discuss a few of their properties. A good intuitive understanding of what a manifold is, as well as what can and cannot be done on a manifold, is essential for differential geometry.

Roughly speaking, a manifold is a space having the "local smoothness structure" of R^n. [R^n, or Euclidean $n-$space, is the set consisting of $n-$tuples, $(x_1, x_2, \ldots, x_n)$, of real numbers.] The idea, then, is to isolate, from all the rich structure on R^n, (e.g., its metric structure, its vector-space structure, its topological structure), that one small bit of structure we call "smoothness." [This process of isolating separate structures for individual attention is, of course, common in mathematics.]

Let M be a set. By an *$n-$chart* on M we understand a subset U of M and a mapping $\varphi : U \to R^n$ having the following two properties: i) φ in one-to-one [That is, no two distinct points of U are taken to the same point of R^n by φ.], and ii) The image of U, that is, the subset $0 = \varphi[U]$ of R^n, is open in R^n. [Recall that a subset 0 of R^n is said to be open if, for any $x \geq 0$, there is a number $\epsilon > 0$ such that, whenever $d(x, y) < \epsilon$, y is also in 0, where $d(x, y)$ is the usual Euclidean distance in R^n.]

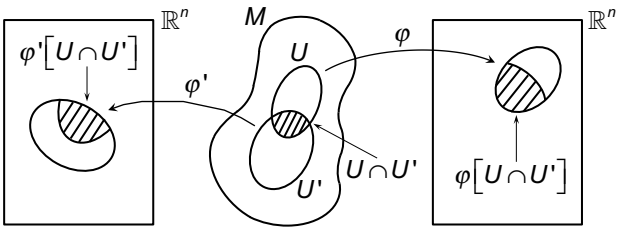


These charts are the mechanism by which we intend to induce our "local smoothness structure" on the set M. They are well-suited for the job. A chart, after all, sets up a correspondence between a set of points U of M and an (open) set of points 0 of R^n. Thus, a chart defines n real-valued functions on U. The points of U can be labeled by the values of these n functions. Since no two distinct points of U have all the same values, such a labeling is unambiguous. (Charts are sometimes called "coordinate patches.") To obtain a manifold, we must lay down enough charts on M, and require that, when two charts overlap (i.e., when their U's overlap in M), the corresponding smoothness structures agree.

Let (U, φ) and (U', φ') be two $n-$charts on the set M. If U and U' intersect in M, then there is induced on their intersection, $U \cap U'$, two smoothness structures. One is obtained from φ [which defines a one-to-one correspondence between $U \cap U'$ and the subset $\varphi[U \cap U']$ of R^n], and the other from φ' [which defines a one-to-one correspondence between $U \cap U'$ and the subset $\varphi'[U \cap U']$ of R^n. By composing these, we obtain a correspondence, defined by $\varphi' \cdot \varphi^{-1}$ and its inverse, $\varphi \cdot \varphi'^{-1}$, between $\varphi[U \cap U']$ and $\varphi'[U \cap U']$ – both subsets of R^n. It is this correspondence which is used to compare the "smoothness structures." The $n-$charts (U, φ) and (U', φ') on M are said to be *compatible* if the following hold: i) $\varphi[U \cap U']$ and $\varphi'[U \cap U']$ are open subsets of R^n. ii) the mappings $\varphi' \cdot \varphi^{-1} : \varphi[U \cap U'] \to \varphi'[U \cap U']$ and $\varphi \cdot \varphi'^{-1} : \varphi'[U \cap U'] \to \varphi[U \cap U']$ are C^∞. [A mapping from a subset of R^n to R^n is just another way of speaking of n functions of n variables. Such a mapping is said to be C^∞ if all the partial derivatives of all n functions exist and are continuous.]

Note that, for compatibility of charts, we only require that they agree in the one structure of interest: smoothness. It is in this way that a single structure is isolated. Note also that (U, φ) and (U', φ') are necessarily compatible if U and U' do not intersect in M.

By an $n-$dimensional *manifold* we shall understand a set M, along with a collection,



$(U_\alpha, \varphi_\alpha)$, of $n-$charts on M (the index α, which labels the various charts, runs over some indexing set A) such that:

M1. Any two charts in the collection are compatible.

M2. The charts cover M, i.e., $\bigcup_\alpha U_\alpha = M$.

M3. Any $n-$chart, which is compatible with all the charts in the collection, is itself in the collection.

M4. If p and p' are distinct points of M, then there exist charts $(U_\alpha, \varphi_\alpha)$ and (U_β, φ_β) such that p is in U_α, p' is in U_β, and such that U_α and U_β do not intersect.

These conditions – or at least the first three of them – are exactly the conditions one would expect. Intuitively, the condition $M1$ states that "if two charts induce a smoothness structure in the same region of M, then those structures must agree." The second condition states that "every part of M has induced on it a smoothness structure." The third condition ensures that we have not placed too much structure on M by cutting down the number of charts. (For example, R^n, which has a much richer structure than that of a manifold, is defined by just "one chart.") Since compatibility looks only at the smoothness structure, and since, by $M3$, we admit all the compatible charts, we conveniently "wash out" all the other structure. The rather ugly fourth condition eliminates certain pathological objects (called non-Hausdorff manifolds) which are not very interesting.

Thus, a manifold consists of a set M along with some charts on that set. We shall usually refer to a manifold as "M," the charts understood. The following sentence is intended as a guide to the literature: What we have called an "$n-$dimensional manifold" might be referred to by others as a "C^∞, Hausdorff, $n-$dimensional manifold, without boundary, which is not necessarily paracompact or connected."

One quickly builds up a sufficiently strong intuition about manifolds that it becomes unnecessary to constantly refer back to conditions $M1 - M4$. The discussion which follows is intended to help in this process.

It might seem that, because of condition $M3$, it would be almost impossible to specify a manifold, for the collection of all those compatible charts will be an extremely large one. The possibility of describing manifolds easily arises from the following fact:

Remark 1. Let M be a set on which there is specified a collection of $n-$charts satisfying $M1$, $M2$, and $M4$. Let C denote the collection of of all $n-$charts which are compatible with those of the given collection. Then the set M, with the charts C, defines a manifold.

Exercise 1: Prove Remark 1. (Everything is clear except condition $M1$, which is not difficult to establish.) Intuitively, Remark 1 states that it is easy to reduce the structure. If you have a set covered by a compatible collection of charts, you can always reduce the structure down to that of a manifold by including more charts.

We give two examples of manifolds.

Example 1. Let M consists of $n-$tuples of real numbers, so, as a point-set, $M = R^n$. We introduce some charts on M: let U be any subset of M which, considered as a subset of R^n, is open, and let φ be the identity mapping from U to R^n. These are certainly charts, and they clearly satisfy $M1$, $M2$, and $M4$. By Remark 1, we obtain a manifold based on M. This manifold is called the manifold R^n.

Example 2. Let M consist of $(n+1)-$tuples of real numbers, $(y_1, \ldots, y_n)$, which satisfy $(y_1)^2 + (y_2)^2 + \ldots + (y_{n+1})^2 = 1$. We introduce some charts on M. Let U_{1^+} consist of points of M with $y_1 > 0$, and let φ_{1^+}, acting on the point $(y_1, \ldots, y_{n+1})$ of U_{1^+} be the point $(y_2, \ldots, y_{n+1})$ of R^n. Let U_{1^-} consist of the points of M with $y_1 < 0$, and let

φ_{1^-}, acting on $(y_1, \ldots, y_{n+1})$ be the point $(y_2, \ldots, y_{n+1})$ of R^n. Let U_{2^+} consist of points of M with $y_2 > 0$, and let φ_{2^+}, acting on the point $(y_1, \ldots, y_{n+1})$ of U_{2^+}, be the point $(y_1, y_3, y_4, \ldots, y_{n+1})$. Continuing in this way, we obtain $(2n + 2)$ charts, (U_{i^+}, φ_{i^+}) and (U_{i^-}, φ_{i^-}), $i = 1, 2, \ldots, (n + 1)$. These $n-$charters are all compatible with each other, and cover M. By including all charts compatible with these, we obtain an $n-$dimensional manifold called the $n-$sphere, S^n.

Exercise 2. Verify that Example 2 leads to a manifold.

There are a number of techniques for obtaining new manifolds from old ones. We describe two such techniques.

Let M and M' be manifolds, of dimensions n and n', respectively. We define a new manifold, $\widetilde{M}$, which is called the *product* of M and M', written $\widetilde{M} = M \times M'$. The dimension of $\widetilde{M}$ will be $(n + n')$. As a point-set, $\widetilde{M}$ is the product of the sets M and M'. That is to say, a point of $\widetilde{M}$ consists of a pair, (p, p'), with p in M and p' in M'. We must introduce some charts on this set M. Let $(U_\alpha, \varphi_\alpha)$ be a chart on M, and $(U_{\alpha'}, \varphi_{\alpha'})$ a chart on M'. Then, since U_α is a subset of M and $U_{\alpha'}$ a subset of M', $\widetilde{U} = U_\alpha \times U_{\alpha'}$ is a subset of $\widetilde{M}$. Now φ_α maps U_α to R^n, while $\varphi_{\alpha'}$ maps $U_{\alpha'}$ to $R^{n'}$. Now let p be a point of U_α, and p' a point of $U_{\alpha'}$, so the pair (p, p') is a point of $\widetilde{U}$. Then $\varphi_\alpha(p)$ is some point, say $(x_1, \ldots, x_n)$, of R^n, while $\varphi_\alpha(p')$ is some point, $(y_1, \ldots, y_{n'})$, of $R^{n'}$. Thus, we can associate with the point (p, p') of $\widetilde{U}$ the point $(x_1, \ldots, x_n, y_1, \ldots, y_{n'})$ of $R^{n+n'}$. In other words, we have just described a mapping, φ, from $\widetilde{U}$ to $R^{n+n'}$. This is a chart on $\widetilde{M}$. Thus, given a chart on M and a chart on M', we obtain a chart on $\widetilde{M}$. It is easy to check that, since M and M' are manifolds, these charts on $\widetilde{M}$ satisfy conditions $M1$, $M2$, and $M4$. (Exercise 3. Do so.) By Remark 1, we have a manifold M.

Example 3. The manifold $R^1 \times S^1$ is called the cylinder. The manifold $S^1 \times S^1$ is called the torus. The manifold $R^n \times R^{n'}$ is the same as the manifold $R^{n+n'}$.

A second method of constructing manifolds is by "cutting holes" in known manifolds. However, only certain types of "holes" can be cut without destroying the manifold structure. Let M be a manifold, and let S be a subset of M. S is said to be *open* in M if, for any point p of S, there is a chart (U, φ), such that p is in U and U is a subset of S. A set is said to be *closed* if its complement is open. (Needless to say, these are the open sets for a topology on M. This topology could also be defined, for example, as the coarsest for which the chart maps are continuous.) Now let C be a closed subset of the manifold M. We define a new manifold $\widetilde{M}$, written $M - C$, which might be called "M with C cut out". As a point-set, $\widetilde{M}$ consists of those points of M which do not lie in C. As charts on $\widetilde{M}$, we take those charts $(U_\alpha, \varphi_\alpha)$ on M for which U_α does not intersect C. Exercise 4. Verify that this M is a manifold, i.e., check conditions $M1 - M4$.

A reasonable fraction of the manifolds which occur in physics can be obtained from the manifold R^n and S^n using the operations of taking products and cutting holes.

2. Smooth Functions: Smooth Mappings

Let M be a manifold, and let f be a (real-valued) function on M. Then, if (U, φ) is a chart on M, f is, in particular, a function on U. Hence, $f \cdot \varphi^{-1}$ is a function on the subset $\varphi[U]$ of R^n. In other words, $f \cdot \varphi^{-1}$ is just a real-valued function of n variables. The function f on M is said to be *smooth* if $f \cdot \varphi^{-1}$ is C^∞ for every chart (U, φ). [Note this technique. To define something on a manifold, one uses the charts to express that something in terms of R^n, where we already know what it means.]

Exercise 5. Let f be a function on M for which $f \cdot \varphi^{-1}$ is C^∞ for some collection of charts satisfying $M2$. Verify that f is a smooth function on M.

Exercise 6. Verify that the function x_1 is smooth on R^n; that the function y_1 is smooth on S^n.

Exercise 7. Let $F(z_1, \ldots, z_k)$ be a C^∞ real-valued function of k real variables. Let $f_1, \ldots, f_k$ be k smooth functions on a manifold M. Show that $F(f_1, \ldots, f_k)$ is a smooth function on a M.

Exercise 8. Let f be a smooth function on M, f' a smooth function on M'. Define a function $\tilde{f}$ on $\widetilde{M} = M \times M'$ as follows: if p is a point of M, p' a point of M', $\tilde{f}(p, p') = f(p) + f'(p')$. Prove that this $\tilde{f}$ is a smooth function on $\widetilde{M}$.

Exercise 9. Let f be a smooth function on M and C a closed subset of M. Prove that $\tilde{f}$, restricted to $\widetilde{M} = M - C$, is a smooth function on $\widetilde{M}$.

We denote by $\mathfrak{F}$ the collection of smooth functions on M.

There is a sense in which the smooth functions on a manifold M characterize the manifold structure of M. This is made more precise by the following:

Remark 2. Let M be a set, and let C and C' be manifold structures on M. (That is, each of C and C' is a collection of charts on M satisfying $M1 - M4$.) Suppose that every smooth functions on (M, C) is also a smooth function on (M, C'). Then $C = C'$.

Exercise 10. Prove Remark 2. (It is perhaps not surprising that one can give axioms on a collection of functions on a set which are necessary and sufficient for those functions to be the smooth functions for some manifold structure on the set.)

Things in mathematics often come in pairs. One introduces some class of objects of interest (e.g., groups, topological spaces, vector spaces, or sets). Then, one introduces certain natural, structure-preserving mappings (sometimes called morphisms) between objects in the same class (e.g., homomorphisms of groups, continuous mappings on topological spaces, linear mappings on vector spaces, or just plain mappings on sets). In the same spirit, manifolds are objects. The corresponding morphisms, which we now define are called differentiable, or smooth mappings.

It is convenient to define a smooth mapping using the smooth functions. [Since the smooth functions "characterize" the manifold structure, it is not surprising that a reasonable definition can be expressed in this way.] Let M and M' be manifolds, and let $\psi : M \to M'$ be a mapping from the point-set M to the point-set M'. If f' is any smooth function on M', then $f' \cdot \psi$ is a function on M. [In words, $f' \cdot \psi$ consists of the following instructions: given a point p of M, take p to M' using ψ, and evaluate f' at the corresponding point, $\psi(p)$ of M'.] The mapping $\psi : M \to M'$ is called a *smooth mapping* if, for every smooth function f' on M', $f' \cdot \psi$ is a smooth function on M.

Exercise 11. Let M be a manifold, and let f be a (real-valued) function on M. Then this f certainly defines a mapping from the manifold M to the manifold R^1. Prove that the function is smooth if and only if the mapping of manifolds is smooth. [Thus, smooth functions can be considered as a special case of smooth mappings.]

Exercise 12. Reexpress the definition of a smooth mapping directly in terms of charts.

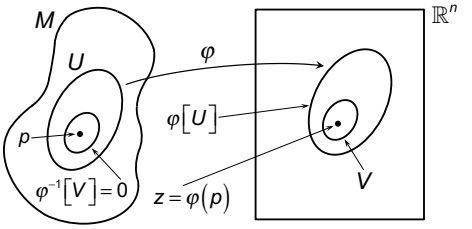
Exercise 13. Let $\widetilde{M} = M \times M'$. Define a mapping $\psi : \widetilde{M} \to M$ as follows. If p is a point of M, and p' is a point of M', so (p, p') is a point of $\widetilde{M}$, then $\psi(p, p') = p$. Prove that ψ is smooth.
Exercise 14. Let $\widetilde{M} = M \times M'$. Define a mapping $\psi : M \to \widetilde{M}$ as follows. Fix a point p' of M'. Then, if p is any point of M, $\psi(p) = (p, p')$. Prove that ψ is smooth.
Exercise 15. Let C be a closed subset of the manifold M. Define a mapping ψ from $M - C$ to M as follows. If p is a point of $M - C$, $\psi(p) = p$. Prove that ψ is smooth.

Morphisms always have the property that the composition of two is another. Smooth mappings are no exception.
Theorem 3. Let M, M', and M'' be manifolds, and let $\psi : M \to M'$ and $\psi' : M' \to M''$ be smooth mappings. Then $\psi' \cdot \psi : M \to M''$ is a smooth mapping.
Proof: Let f'' be a smooth function on M''. We must show that $f'' \cdot \psi' \cdot \psi$, is a smooth function on M. Since f'' is smooth on M'' and ψ' is a smooth mapping, $f'' \cdot \psi'$ is a smooth function on M'. But, since $f'' \cdot \psi'$ is a smooth function on M' and ψ is a smooth mapping, $f'' \cdot \psi' \cdot \psi$ is a smooth function on M.
Exercise 16. What are the compositions of the mappings in Exercises 13 and 14?

An important concept in mathematics is that of an isomorphism between objects. Roughly speaking, two objects are isomorphic if they are identical as far as the structure under consideration is concerned. [For groups and vector spaces, isomorphisms are called isomorphisms; for topological spaces, homeomorphisms; for sets, bijections.] The "isomorphisms" for manifolds are called differeomorphisms. Let M and M' be manifolds, and let $\psi : M \to M'$ be a smooth mapping. This ψ is said to be a *diffeomorphism* if ψ is one-to-one and onto (so that ψ^{-1} is well-defined), and $\psi^{-1} : M' \to M$ is also a smooth mapping. Given two manifolds, M and M', there may or may not exist a diffeomorphism between them. If there does exist one, M and M' are said to be *diffeomorphic*. One occasionally writes $M = M'$ to indicate that M and M' are diffeomorphic.
Example 4. Define a mapping $\psi : S^n \to R^n$ as follows. If $(y_1, \ldots, y_{n+1})$ is any point of S^n, $\psi(y_1, \ldots, y_{n+1})$ is the point $(y_1, \ldots, y_n)$ of R^n. Although this ψ is smooth, it is not one-to-one: for example, $\psi(0, \ldots, 0, -1) = \psi(0, \ldots, 0, 1)$. Hence, ψ is not a diffeomorphism.
Example 5. Let $\psi : R^1 \to R^1$ be the smooth mapping $\psi(x) = x^3$. Since this ψ is one-to-one and onto its inverse, $\psi^{-1}(x) = x^{1/3}$, exists. But, since the inverse is not smooth, ψ is not a diffeomorphism.
Example 6. Let C be the closed subset of R^2 consisting of the single point $(0, 0)$. Define a mapping ψ from $R^2 - C$ to $R^2 - C$ as follows: $\psi(x_1, x_2) = (r^{-2}x_1, r^{-2}x_2)$, where $r^2 = (x_1)^2 + (x_2)^2$. This ψ is clearly a diffeomorphism.
Exercise 17. Verify that $S^n - C$, where C is the point $(0, \ldots, 0, 1)$ of S^n, is diffeomorphic with R^n. Verify that $R^n - C$, where C is the point $(0, \ldots, 0)$ of R^n, is diffeomorphic with $S^{n-1} \times R^1$.
Exercise 18. Prove that the composition of two diffeomorphisms is a diffeomorphism. Prove that the inverse of a diffeomorphism is a diffeomorphism.
Exercise 19. Prove that, if two manifolds are diffeomorphic, then they have the same dimension.

As a final example to illustrate the structure possessed by a manifold, we establish a theorem to the effect that manifolds are "locally homogeneous." [The theorem is somewhat analogous to the following fact about vector spaces: If v and w are non zero elements of a vector space V, then there is an isomorphism from V to V which takes v to w.] Thus, manifolds, like vector spaces, are barren, devoid of landmarks.
Theorem 4. Let p be a point of the manifold M. Then there is an open set O, containing p, with the following property: given any point p' in O, there exists a diffeomorphism $\psi : M \to M$ with $\psi(p) = p'$. Proof: Let (U, φ) be a chart such that p is in U. Let $\varphi(p) = z = (z_1, \cdots, z_n)$, a point of R^n. Choose $\epsilon > 0$ such that, whenever x is a point of R^n with $d(x, z) < \epsilon$, x is in $\varphi[U]$. Denote by V the collection of all x in R^n with $d(x, z) < \epsilon$. The subset $O = \varphi^{-1}[V]$ of M is our candidate for the O in the theorem. [Exercise 20.

Why is O open in M?]

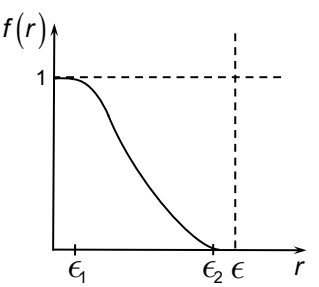


Let p' be a point of O, so $\varphi(p') = z' = (z'_1, \cdots, z'_n)$, a point of R^n, is in V. Set $\epsilon' = d(z, z') < \epsilon$. Define a mapping Λ from V to V as follows:

$\Lambda(x_1, \cdots, x_n) = (x_1, \cdots, x_n) + f(r)(z'_1 - z_1, \cdots, z'_n - z_n)$,

where we have set $r = [d(x, z)]$, and where $f(r)$ is a function of one variable with the following properties: i) $f(r)$ is C^∞, ii) there is an $\epsilon_1 > 0$ such that $f(r) = 1$ for $r < \epsilon_1$, iii) there is an $\epsilon_2 < \epsilon$ such that $f(r) = 0$ for $r > \epsilon_2$, and iv) $|df/dr| < (\epsilon')^{-1}$. [Since $\epsilon < \epsilon$, there exists such an $f(r)$.] These conditions ensure that Λ is a smooth mapping from V to V (conditions i and ii), that $\Lambda(z) = z'$ (condition ii), that Λ is the identity near the edge of V (condition iii), and that Λ^{-1} exists and is smooth (conditions i, ii, iii, and iv). [Exercise 21. Verify these properties.]

We now define the mapping ψ from M to M. Let q be a point of M. If q is not in O, set $\psi(q) = q$. If q is in O, set $\psi(q) = \varphi^{-1} \cdot \Lambda \cdot \varphi(q)$. It is clear that this ψ is smooth, one-to-one, onto, and that $\psi(p) = p'$. [Exercise 22. Check that ψ is a diffeomorphism.]



Intuitively, since we can shift points a little bit in R^n, preserving the smoothness structure (a fact made explicit by the expression, above, for $\Lambda(x)$), and since M has the local smoothness structure of R^n, we can shift points a little in M, preserving the smoothness structure.

The statement of Theorem 4 is interesting because of what it says about manifolds. The proof of Theorem 4 is interesting because it illustrates a useful technique: certain local properties of R^n can, using the charts, be pulled back to properties of manifolds.

Exercise 23. State and prove: S^n and R^n are globally homogeneous.

3. Vectors at a point

We now have the arena for differential geometry: a manifold. The natural things which exist in such an environment are objects called tensor fields. These tensor fields will be defined in the next few sections. Although the fields are by far the most important, it is convenient to first introduce tensors at a point.

We begin with vectors at a point, $\underline{x} = (\underline{x}^1, \ldots, \underline{x}^n)$ of R^n. A vector in R^n at the point $\underline{x}$ can be represented by its components, $(\xi^1, \ldots, \xi^n)$, with respect to the axes in R^n. That is to say, a vector in R^n at $\underline{x}$ is equivalent to n real numbers. However, while this representation of vectors is natural in R^n, it is not a very convenient one for discussing vectors on manifolds (in which there are no natural "axes"). Denote by $\mathfrak{F}(R^n)$ the collection of all smooth functions on R^n so an element of $\mathfrak{F}(R^n)$ consists of a C^∞, real-valued function, $f(x^1, \ldots, x^n)$, of n real variables. If $(\xi^1, \ldots, \xi^n)$ is the components of any vector at $\underline{x}$, and f is any smooth function on R^n, we set

$$\xi(f) = \xi^1 \left.\frac{\partial f}{\partial x^1}\right|_{\underline{x}} + \cdots + \xi^n \left.\frac{\partial f}{\partial x^n}\right|_{\underline{x}} \tag{1}$$

This $\xi(f)$, which is just a real number, is called the directional derivative of f in the direction of the vector represented by $(\xi^1, \ldots, \xi^n)$. It follows immediately from the elementary properties of the derivative that $\xi(f)$ satisfies the following conditions:

DD1. $\xi(f + g) = \xi(f) + \xi(g)$.
DD2. $\xi(fg) = f(\underline{x})\xi(g) + g(\underline{x})\xi(f)$.
DD3. If $f = \text{const.}$, then $\xi(f) = 0$.

Thus, every vector in R^n at $\underline{x}$ defines, via (1), a mapping $\xi(f)$ from $\mathfrak{F}(R^n)$ to R, satisfying DD1, DD2, and DD3. We have :

Proposition 5. Eqn. (1) defines a one-to-one correspondence between $n-$tuples, $(\xi^1, \ldots, \xi^n)$ and mappings from $\mathfrak{F}(R^n)$ to R satisfying DD1, DD2, and DD3.

Proof: Two things must be shown: i) if $(\xi^1, \ldots, \xi^n)$ and $(\eta^1, \ldots, \eta^n)$ have the property that $\xi(f) = \eta(f)$ for all f in $\mathfrak{F}(R^n)$ (with $\xi(f)$ and $\eta(f)$ defined by Eqn. (1)), then $\xi^1 = \eta^1, \ldots, \xi^n = \eta^n$, and ii) if $\mu(f)$ is a mapping from $\mathfrak{F}(R^n)$ to R, satisfying DD1, DD2, and DD3, then $\mu(f)$ is of the form (1) for some $(\mu^1, \ldots, \mu^n)$.

i) Let $\xi(f) = \eta(f)$ for all f. Settings $f = x^1$, we have, from Eqn. (1), $\xi' = \xi(x^1) = \eta(x^1) = \eta'$. Similarly for $x^2, \ldots, x^n$.

ii) Let $\mu(f)$ satisfy DD1, DD2, and DD3. Define n numbers, $\mu^1 = \mu(x^1), \ldots, \mu^n = \mu(x^n)$. We shall show that, with this $(\mu^1, \ldots, \mu^n)$, Eqn. (1) holds for all f. Let f be in $\mathfrak{F}(R^n)$, and write f in the form

$$f(x) = f(\underline{x}) + (x^1 - \underline{x}^1)g_1(x) + \cdots + (x^n - \underline{x}^n)g_n(x) \tag{2}$$

with $g_1(x), \ldots, g_n(x)$ in $\mathfrak{F}(R^n)$. [See Lemma 6 below.] Then, from DD1 and DD2,

$$\begin{aligned}\mu(f) = \mu(f(\underline{x})) + (x^1 - \underline{x}^1)|_{\underline{x}}\, \mu(g_1) + g_1(x)|_{\underline{x}}\, \mu(x^1 - \underline{x}^1) \\ + \cdots + (x^n - \underline{x}^n)|_{\underline{x}}\, \mu(g_n) + g_n(x)|_{\underline{x}}\, \mu(x^n - \underline{x}^n)\end{aligned} \tag{3}$$

By DD3, $\mu(f(\underline{x})) = 0$; by DD1 and DD3, $\mu(x^i - \underline{x}^i) = \mu^i$; by Eqn.(2), $g_i|_{\underline{x}} = \partial f/\partial x^i|_{\underline{x}}$; clearly, $(x^i - \underline{x}^i)|_{\underline{x}} = 0$ $(i = 1, 2, \ldots, n)$. Eqn. (1) now follows from Eqn. (3).

In the proof of Proposition 5, we made use of the following fact:

Lemma 6. If f is in $\mathfrak{F}(R^n)$, then f can be written in the form (2) (where $\underline{x}$ is some fixed point of R^n) for some $g_1(x), \ldots, g_n(x)$ in $\mathfrak{F}(R^n)$.

Proof:

$$f(x) - f(\underline{x}) = \int_0^1 \frac{\partial}{\partial t} f(\underline{x} + t(x - \underline{x}))dt$$
$$= (x^1 - \underline{x}^1) \int_0^1 \frac{\partial}{\partial x^1} f(\underline{x} + t(x - \underline{x}))dt$$
$$+ \cdots + (x^n - \underline{x}^n) \int_0^1 \frac{\partial}{\partial x^n} f(\underline{x} + t(x - \underline{x}))dt$$

Proposition 5 provides a different, and more fruitful, representation of vector in R^n.

With the remarks above as motivation, we now return to manifolds. Let M be a manifold, and, as in Sect. 2, let $\mathfrak{F}$ be the collection of smooth functions on M. Then (e.g., by Exercise 7), the (pointwise) sum and product of elements of $\mathfrak{F}$ is in $\mathfrak{F}$. [This $\mathfrak{F}$ has the structure of a ring.] Fix a point p of M. By a *contravariant vector* in M at p, we mean a mapping $\xi : \mathfrak{F} \to R$ satisfying DD1, DD2, and DD3 (with, of course, the $\underline{x}$ in DD2 replaced by p). The collection of all contravariant vectors in M at p will be written $\mathfrak{F}_\cdot(p)$. If ξ and η are in $\mathfrak{F}_\cdot(p)$, m is a real number, and f is in $\mathfrak{F}$, set $(\xi + m\eta)(f) = \xi(f) + m\eta(f)$. With this as the definition of the sum of two elements of $\mathfrak{F}_\cdot(p)$, and of the product of an element of $\mathfrak{F}_\cdot(p)$ with a real number, $\mathfrak{F}_\cdot(p)$ clearly becomes a vector space.

The contravariant vectors can be represented more explicitly in terms of charts. Let (U, φ) be a chart on M. Then, if f is a smooth function on M, $f \cdot \varphi^{-1} = \tilde{f}(x^1, \ldots, x^n)$ is a smooth function on $\varphi[U]$. Now suppose that the point p is in U, and set $\underline{x} = \varphi(p)$. Choose $h^1, \ldots, h^n$ in $\mathfrak{F}$ such that, in some open subset of R^n containing $\underline{x}$, $\tilde{h}^1 = (x^1 - \underline{x}^1), \ldots, \tilde{h}^n = (x^n - \underline{x}^n)$. [Exercise 24. Why do such h's exist?] If ξ is a contravariant vector in M at p, the numbers $\xi' = \xi(h^1), \ldots, \xi^n = \xi(h^n)$ are called *components* of ξ with respect to the chart (U, φ).

Theorem 7. The components of ξ are independent of the choice of the h's. Furthermore, given n numbers, $(\xi', \ldots, \xi^n)$, there exists precisely one contravariant vector in M at p with components $(\xi', \ldots, \xi^n)$,

Proof: Let f be in $\mathfrak{F}$, and set

$$f = f(p) + g_1 h^1 + \cdots + g_n h^n + s^2 t \tag{4}$$

where $g_1, \ldots, g_n, s, t$ are in $\mathfrak{F}$ and $s(p) = 0$. [To construct such an expression, proceed as follows. First choose $g_1, \ldots, g_n$ so $f = f(p) + g_1 h^1 + \ldots + g_n h^n$ in some open set containing p. Then choose any s which is positive except at p, where it vanishes. Finally, choose t so that (4) is satisfied.] Using the argument surrounding Eqn. (3), and the fact that $\xi(s^2 t) = 2s(p)t(p)\xi(s) + [s(p)]^2 \xi(t) = 0$, we obtain

$$\xi(f) = \xi^1 \frac{\partial \tilde{f}}{\partial x^1}\bigg|_{\underline{x}} + \cdots + \xi^n \frac{\partial \tilde{f}}{\partial x^n}\bigg|_{\underline{x}} \tag{5}$$

That the components are independent of the h's is clear from Eqn. (5). To prove the second statement, observe that, if $(\xi^1, \ldots, \xi^n)$ are n numbers, then Eqn. (5) defines a contravariant vector in M at p with components $(\xi^1, \ldots, \xi^n)$. Thus, if M is an n-dimensional manifold, and p is a point of M, then $\mathfrak{F} \cdot (p)$ is an n-dimensional vector space.

The components assigned to a vector depend, of course, on the choice of chart. We derive the well-known formula for this dependence. Let (U, φ) and (U', φ') be two charts, both containing the point p of M. Then $U\hat{U}'$ is coordinated in two ways, once by φ and once by φ' . That is to say, we have a smooth, one-to-one, onto mapping, $\varphi' \cdot \varphi^{-1}$, from $\varphi[U\hat{U}']$ to $\varphi'[U\hat{U}']$, both subsets of R^n. In other words, we have a functions of n variables, $x'^1(x^1, \ldots, x^n), \ldots, x'^n(x^1, \ldots, x^n)$. Let f be a smooth function on M. Then, setting $\tilde{f} = f \cdot \varphi^{-1}$, $\tilde{f}' = f \cdot \varphi'^{-1}$, we have $f(x^1, \ldots, x^n) = f'(x'^1(x^1, \ldots, x^n), \ldots, x'^n(x^1, \ldots, x^n))$. Now let ξ be a contravariant vector in M at p and let $(\xi^1, \ldots, \xi^n)$ and $(\xi'^1, \ldots, \xi'^n)$ be the

components of ξ with respect to (U, φ) and (U', φ'), respectively. It follows from Eqn. (5) that

$$\begin{aligned}\sum_{i=1}^{n} \xi'^{i} \frac{\partial \tilde{f}'}{\partial x'^{i}}\Big|_{\underline{x}'} = \xi(f) &= \sum_{i=1}^{n} \xi^{i} \Big[\frac{\partial \tilde{f}}{\partial x^{i}}\Big|_{\underline{x}} \Big] \\ &= \sum_{i=1}^{n} \xi^{i} \Big[\sum_{j=1}^{n} \frac{\partial \tilde{f}'}{\partial x'^{j}}\Big|_{\underline{x}'} \frac{\partial x'^{j}}{\partial x^{i}}\Big|_{\underline{x}} \Big]\end{aligned} \tag{6}$$

where $\underline{x} = \varphi(p)$ and $\underline{x}' = \varphi'(p)$. But Eqn. (6) must hold for all smooth f. Therefore,

$$\xi'^{1} = \sum_{i=1}^{n} \xi^{i} \frac{\partial x'^{1}}{\partial x^{i}}\Big|_{\underline{x}}, \ldots, \quad \xi'^{n} = \sum_{i=1}^{n} \xi^{i} \frac{\partial x'^{n}}{\partial x^{i}}\Big|_{\underline{x}} \tag{7}$$

[Exercise 25. Why does (7) follow from (6)?] Eqn. (7) describes the behavior of components (of a fixed vector) under change in choice of chart.

Exercise 26. Let p be a point of a manifold M. Suppose we are given a mapping from charts on M containing p, to $n-$tuples of real numbers, subject to (7). Show that such a mapping defines a contravariant vector in M at p.

Exercise 27. Show that, given a nonzero vector, ξ at p and an n-tuple of real numbers, $(\xi^1, \ldots, \xi^n)$, not all zero, there exists a chart with respect to which the components of ξ are $(\xi^1, \ldots, \xi^n)$.

4. Linear Algebra

In this section, we shall review some basic facts about linear algebra, and introduce some notation.

Let , V^a, $V^b, \ldots$ be finite-dimensional vector spaces (over the reals). It is convenient to indicate to which V a vector belongs by means of superscript. Thus, an element of V^a would be written ξ^a, an element of V^c as η^c, etc. Hence, addition of two vectors is defined when and only when those vectors have the same superscript.

We define the dual of a vector space. Let $\mu : V^a \to R$ be linear, i.e., let μ be such that $\mu(\xi^a + m\eta^a) = \mu(\xi^a) + m\mu(\eta^a)$, for any number m and any ξ^a and η^a in V^a. If μ and ν are two such linear maps, and m is a real number, we define a new linear map, $(\mu + m\nu)$, by $(\mu + m\nu)(\xi^a) = \mu(\xi^a) + m\nu(\xi^a)$. With this definition of addition of linear maps, and multiplication of linear maps by (real) numbers, the linear maps from V^a to R form a vector space, called the dual of V^a. The dual of V^a will be written V_a, and membership in V_a will be indicated by the subscript "a". Thus, the μ above would be written μ_a. Finally, instead of $\mu(\xi^a)$, we write $\mu_a\xi^a$ or $\xi^a\mu_a$. It is clear that formulae such as

$$(\mu_a + m\nu_a)(\xi^a + k\eta^a) = \mu_a\xi^a + m\nu_a\xi^a + k\mu_a\eta^a + mk\nu_a\eta^a$$

are true. Similarly, the dual of V^b will be written V_b, and membership in V_b indicated by the subscript "b", etc. Thus, λ_c would represent an element of V_c, i.e., a linear mapping from V^c to R. In order that the sum of two vectors be defined, it is necessary that i) both vectors have subscripts or both vectors have superscripts, and ii) the subscripts or superscripts be the same letter.

We next define multilinear maps. Choose any finite (ordered) list of vector spaces, from the collection $V^a, V_a, V^b, V_b, \ldots$, having the following property: no vector space appears more than ones in the list, and no vector space appears together with its dual. (That is, no letter appears more than ones in the list.) A typical such list is (V^c, V_a, V_e, V^d). A mapping α from $V^c \times V_a \times V_e \times V^d$ (a product of sets) to R is said to be multilinear if α is linear in each variable separately, i.e., if

$$\alpha(\xi^c + m\eta^c, \mu_a, \lambda_e, \tau^d) = \alpha(\xi^c, \mu_a, \lambda_e, \tau^d) + m\alpha(\eta^c, \mu_a, \lambda_e, \tau^d),$$

$$\alpha(\xi^c, \mu_a + m\nu_a, \lambda_e, \tau^d) = \alpha(\xi^c, \mu_a, \lambda_e, \tau^d) + m\alpha(\xi^c, \nu_a, \lambda_e, \tau^a), \text{ etc.}$$

The collection of all such multilinear maps will be denoted by $V_c{}^{ae}{}_d$, and an element of $V_c{}^{ae}{}_d$ (e.g., α above) will be written $\alpha_c{}^{ae}{}_d$. Instead of $\alpha(\xi^c, \mu_a, \lambda_e, \tau^d)$, we write $\alpha_c{}^{ae}{}_d\xi^c\mu_a\lambda_e\tau^d$. Defining the sum of two multilinear mappings, and the product of a multilinear mapping with a number, by $(\alpha_c{}^{ae}{}_d + m\beta_c{}^{ae}{}_d)(\xi^c\mu_a\lambda_e\tau^d) = \alpha_c{}^{ae}{}_d\xi^c\mu_a\lambda_e\tau^d + m\beta_c{}^{ae}{}_d\xi^c\mu_a\lambda_e\tau^d$, we see that $V_c{}^{ae}{}_d$ has the structure of a vector space.

Thus, starting with the vector space $V^a, V^b, \ldots$, we obtain a much larger collection of vector spaces, e.g., $V^{am}{}_{b\ vk}^{\ c}$, etc. The elements of these vector spaces are called *tensors*. The superscripts attached to a tensor are called *contravariant indices*, the subscripts *covariant indices*. The *rank* of a tensor is defined as the total number of indices it possesses. Tensors of rank one are usually called *vectors*. Real numbers, when considered as tensors of rank zero, are called *scalars*.

What operations are available between these tensors? Of course, one can multiply a tensor by a scalar. This operation can be generalized. Consider two tensors which have no index letter in common (i.e., either "a" appears in precisely one of the tensors, or "a" appears in nether, etc. for "b", "c", etc....), e.g., consider $\alpha_m{}^e$ and $\beta^{ca}{}_b{}^p$. Then the expression $(\alpha_m{}^e\xi^m\eta_e)(\beta^{ca}{}_b{}^p\mu_c\nu_a\sigma^b\tau_p)$ evidently defines a multilinear map from $V^m \times V_e \times V_c \times V_a \times V^b \times V_p$ to R. That is, we have defined an element of $V_m{}^{eca}{}_b{}^p$. This element will be written $\alpha_m{}^e\beta^{ca}{}_b{}^p$. Thus, two tensors having the property that they share no index letter

define a third tensor. This operation is called *outer product*. The rank of the outer product of two tensors is the sum of the ranks of the two tensors. Multiplication of a tensor by a scalar can be considered as a special case of the outer product.

Exercise 28. Check that $\alpha_n{}^e(\beta^{ca}\gamma_n{}^{pq}) = (\alpha_m{}^e\beta^{ca})\gamma_n{}^{pq}$. (Hence, parenthesis in such expressions are unnecessary.)

The second operation is addition. Of course, one can add two tensors having exactly the same index structure (i.e., having the same index letters, in the same order, and in the same locations), e.g., $\alpha_a{}^r{}_c + \beta_a{}^r{}_c$. It is convenient, however, to permit addition in more general situations. Evidently, the multilinear maps from $V^c \times V_n \times V_p$ to R are in one-to-one correspondence with multilinear maps from $V_n \times V_p \times V^c$ to R etc. That is to say, there are natural isomorphisms between the vector spaces $V_c{}^{np}$, $V^{np}{}_c$,$V^n{}_c{}^p$, $V^p{}_c{}^n$, $V^{pn}{}_c$, $V_c{}^{pn}$. Thus, an element of one of these vector spaces defines an element of each space. Whenever addition or equality is indicated between two tensors in isomorphic vector spaces, it is understood that one tensor is to be taken, using the isomorphism, into the vector space of the other, and there the sum is to be evaluated, or equality is to hold.

Example 7. The equation $\gamma^{np}{}_c = \alpha^n{}_c{}^p + \beta_c{}^{pn}$ makes the following statement. Using the isomorphism between $V_c{}^{pn}$ and $V^n{}_c{}^p$, take $\beta_c{}^{pn}$ (an element of $V_c{}^{pn}$) to $V^n{}_c{}^p$. Add the result to $\alpha^n{}_c{}^p$ (an element of $V^n{}_c{}^p$). Using the isomorphism between $V^n{}_c{}^p$ and $V^{np}{}_c$, take the resulting sum (an element of $V^n{}_c{}^p$) to $V^{np}{}_c$. The result is the same as the element $\gamma^{np}{}_c$ of $V^{np}{}_c$. Thus, addition is defined between tensors having the property that they have exactly the same letters as contravariant indices and exactly the same letters as covariant indices.

Exercise 29. State and prove: $\xi^a{}_p\eta^{dm} = \eta^{dm}\xi^a{}_p$.

In fact, we have developed the formalism above in a context slightly more general that we shall need. Let V be a single finite-dimensional vector space, and let V^a, V^b, ..., be copies of V. (That is to say, we are given isomorphisms between V^a and V, between V^b and V, etc.) Using the construction above, we obtain in this case, just as in the more general case, tensors. The same operations as above apply: the outer product is defined between any two tensors having no index letter in common: addition and equality are defined between any two tensors having no index letter in common; addition and equality are defined between any two tensors having exactly the same contravariant index letters and exactly the same covariant index letters. However, two additional operations are available in this special case.

Consider one of our vector space, e.g., $V^m{}_{ap}{}^d$. Choose two index letters, the first of which appears in $V^m{}_{ap}{}^d$, and the second of which does not, e.g., "p"and "b". The isomorphism between V^p to V^b then, evidently, defines an isomorphism between $V^m{}_{ap}{}^d$ and $V^m{}_{ab}{}^d$. Thus, any element, e.g., $\alpha^m{}_{ap}{}^d$, of $V^m{}_{ap}{}^d$ defines some element of $V^m{}_{ab}{}^d$. This element of $V^m{}_{ab}{}^d$ will be written $\alpha^m{}_{ab}{}^d$. We say that $\alpha^m{}_{ab}{}^d$ has been obtained from $\alpha^m{}_{ap}{}^d$ by *index substitution*. Tensors, which are obtained from each other by index substitution are not written as equal, i.e., we do not write $\alpha^m{}_{ab}{}^d = \alpha^m{}_{ap}{}^d$. Index substitution is always indicated as above: one changes the letter which appears as an index, without changing the location of that index or the base letter. ("α" in the example above).

Example 8. Let $\beta_{mc}{}^d = \tau^d{}_m\mu_c$. Perform index substitution on $\beta_{mc}{}^d$ to obtain $\beta_{mc}{}^a$. Perform index substitution on $\tau^d{}_m$ to obtain $\tau^a{}_m$. Then $\beta_{mc}{}^a = \tau^a{}_m\mu_c$.

Example 9. Let β^{ab} be a tensor. Perform index substitution on β^{ab} to obtain β^{mb}; perform index substitution on β^{mb} to obtain β^{mn}; perform index substitution on β^{mn} to obtain β^{bn}; perform index substitution on β^{bn} to obtain β^{na}. Then, in general, it will not be true that $\beta^{ab} = \beta^{ba}$. If, however, ξ^a is a vector, ξ^b is the result of index substitution on ξ^a, and $\beta^{ab} = \xi^a\xi^b$, then it is true that $\beta^{ab} = \beta^{ba}$.

The second operation is called contraction. Consider a tensor with at least one covariant index and at least one contravariant index, e.g., $\alpha^{pbr}{}_c$. Choose two indices, one

covariant and one contravariant, of $\alpha^{pbr}{}_{c}$, e.g.,"b" and "c". This $\alpha^{pbr}{}_{c}$ can always be written as a sum of outer products of the form

$$\alpha^{pbr}{}_{c} = \lambda^{pr}\xi^{b}\eta_{c} + \cdots + \rho^{pr}\mu^{b}\nu_{c}$$

Now consider the tensor

$$\lambda^{pr}(\xi^{b}\eta_{b}) + \cdots + \rho^{pr}(\mu^{b}\nu_{b})$$

an element of V^{pr} (the quantities $\xi^{b}\eta_{b}$, etc. are just scalars). This tensor (which is uniquely determined by $\alpha^{pbr}{}_{c}$, and the choice of indices "b" and "c") is called the *contraction* of $\alpha^{pbr}{}_{c}$ over the indices "b" and "c". It is written $\alpha^{pbr}{}_{b}$, or $\alpha^{pmr}{}_{m}$, etc., where the repeated index is any letter which does not appear elsewhere in α. When several contractions are applied to the same tensor, one uses different repeated indices to avoid confusion. Thus, contraction reduces the rank of a tensor by two. Simultaneous outer product and contraction are easily indicated by the notation, e.g., if $\alpha^{mb}{}_{ad} = \lambda^{m}{}_{d}\xi_{a}{}^{b}$, then $\alpha^{mc}{}_{ac}$ can be written $\lambda^{m}{}_{c}\xi_{a}{}^{c}$. The notations for the action of the dual (i.e., $\mu_{a}\xi^{a}$) and the action of a multilinear map (i.e., $\alpha^{mb}{}_{a}\xi_{m}\eta_{b}\lambda^{a}$) are clearly consistent with the notation for contraction, and, in fact we shall consider these actions as special cases of the contraction operation.

To summarize, given a single finite-dimensional vector space V, we obtain a collection of tensors, indexed by letters. On these tensors, there are defined four operations: i) outer product, defined between two tensors having no index letter in common, ii) addition, defined between two tensors having exactly the same contravariant index letters and exactly the same covariant index letters, iii) index substitution, which results in replacing one index letter of a tensor with any other letter which does appear elsewhere in the tensor, and iv) contraction, which results in replacing a pair of index letters, one contravariant and one covariant, by the same index letter which does not appear elsewhere in the tensor.

This notation is due to Penrose. Although the rules may sound rather formidable at first, one quickly gets used to them, and automatically performs the various operations where appropriate. The notation represents an important labor-saving device.

Example 10. The study of linear operators on a finite-dimensional vector space V is the study of tensors $\alpha^{a}{}_{b}$ over V. Composition of operators is defined by $\alpha^{a}{}_{m}\beta^{m}{}_{b}$, and the trace of an operator is $\alpha^{a}{}_{a}$.

Example 11. Let $\alpha^{ab} = \beta^{ab} + \beta^{ba}$. Then $\alpha^{ab} = \alpha^{ba}$. If $\gamma_{mn} = -\gamma_{nm}$, then $\alpha^{ab}\gamma_{ab} = 0$.

Example 12. If $\alpha^{ab} = \alpha^{ba}$. Then, for any ξ_{p}, $\xi_{b}\alpha^{ab} = \xi_{d}\alpha^{da}$.

Example 13. The following is meaningless: $\xi^{am}{}_{bc}{}^{da}$.

Example 14. If $\tau_{b}{}^{da} = \xi^{a}\mu_{b}{}^{d} + \eta^{d}\lambda_{b}{}^{a}$, then $\tau_{m}{}^{dm} = \xi^{c}\mu_{c}{}^{d} + \eta^{d}\lambda_{e}{}^{e}$

Example 15. Let $\delta^{a}{}_{b}$ be the (unique) element of $V^{a}{}_{b}$ having the property $\delta^{a}{}_{b}\xi^{b} = \xi^{a}$ for every ξ^{b}. Then $\delta^{a}{}_{a}$ = dimension of V. Furthermore, $\delta^{a}{}_{c}\alpha^{mb}{}_{ad} = \alpha^{mb}{}_{cd}$.

5. Tensors at a point

We have seen in Sect. 3 that the collection, $\mathfrak{F}'(p)$, of contravariant vectors at the point p of an $n-$dimensional manifold M form an $n-$dimensional vector space. We have seen in Sect. 4 that, given an $n-$dimensional vector space, one can construct tensors, e.g., $\lambda^p{}_{mcs}{}^b$, over that vector space. The tensors which result when this construction is applied to the vector space $\mathfrak{F}'(p)$ are called *tensors* in M at the point p. Between tensors in M at p, we have outer product, addition, index substitution, and contraction.

We illustrate by returning again to components. Let (U, φ) be a chart containing p, and let η_a be a covariant vector at p. Set $\eta_1 = \eta_a \xi^a$, where ξ^a is the contravariant vector at p with components $(1, 0 \ldots, 0)$; set $\eta_1 = \eta_a \mu^a$, where μ^a is the contravariant vector at p with components $(0, 1, 0, \ldots, 0)$, etc. The n numbers $(\eta_1, \eta_2, \ldots, \eta_b)$ are called the *components* of η_a with respect to the chart (U, φ). Evidently, if ξ^a is a contravariant vector at p with components $(\xi^1, \xi^2, \ldots, \xi^n)$, then $\eta_a \xi^a = \eta_1 \xi^1 + \cdots + \eta_n \xi^n$.

Now let (U', φ') be a second chart containing p. Let $\underline{x} = \varphi(p)$, $\underline{x}' = \varphi'(p)$, and let $x^1(x'^1, \ldots, x'^n), \ldots, x^n(x'^1, \ldots, x'^n)$ be the n functions of n variables defined by $\varphi \cdot \varphi'^{-1}$. Then, evidently, $\sum_{k=1}^n \frac{\partial x^i}{\partial x'^k}|_{\underline{x}'} \frac{\partial x'^k}{\partial x^j}|_{\underline{x}} = (1$ if $i = j, 0$, otherwise: $i, j, = 1, \ldots, n)$. Denote by $(\eta_1 \ldots, \eta_n)$ and $(\eta'_1 \ldots, \eta'_n)$ the components of η_a with respect to (U, φ) and (U, φ') respectively. Then

$$\sum_{i=1}^n \eta_i \xi^i = \eta_a \xi^a = \sum_{j=1}^n \eta'_j \xi'^j = \sum_{j=1}^n \sum_{i=1}^n \eta'_j \left. \frac{\partial x'^j}{\partial x^i} \right|_{\underline{x}} \xi^i$$

where we have used Eqn. (7). Evidently, this equation can hold for all ξ^a only if $\eta'_i = \sum_{j=1}^n \eta_j \frac{\partial x^j}{\partial x'^i}|_{\underline{x}'}$, $(i = 1, \ldots, n)$. This is the transformation law for the components of a (fixed) covariant vector.

Exercise 30. Define the components of an arbitrary tensor, e.g., $\alpha^a{}_{bc}$. Prove that these components have the following transformation law:

$$\alpha'^i{}_{jk} = \sum_{l=1}^n \sum_{p=1}^n \sum_{q=1}^n \alpha^l{}_{pq} \left. \frac{\partial x'^i}{\partial x^l} \right|_{\underline{x}} \left. \frac{\partial x^p}{\partial x'^j} \right|_{\underline{x}'} \left. \frac{\partial x^q}{\partial x'^k} \right|_{\underline{x}'}$$

$(i, j, k = 1, \ldots, n)$.

6. Tensor Fields

We now introduce the main objects of interest on a manifold: tensor fields.

Let M be a manifold. A *tensor field* on M consists of an assignment, to each point of M, a tensor at that point, where all the tensors at various points have the same index structure (i.e., the same index letters in the same locations). Thus, if $\alpha_m{}^{acn}{}_p$ is a tensor field on M, then, for each point p of M, $\alpha_m{}^{acn}{}_p(p)$ is a tensor at the point p.

Example 16. whereas the components of a tensor at a point consist of an array of real numbers, the components of a tensor field are functions of the n coordinates, $x^1, \ldots, x^n$

Evidently, the four tensor operations – addition, outer product, index substitution, and contraction – can be performed pointwise. Hence, these operations are also defined on tensor fields.

Example 17. Let $\alpha_m{}^{acn}{}_p$ be a tensor field. Then, for each point p $\alpha_m{}^{acn}{}_p(p)$ is a tensor at p. Since contraction is defined for tensors at p, we have, for each point p, a tensor $\alpha_d{}^{adn}{}_p(p)$. Thus, we have a tensor field, written $\alpha_d{}^{adn}{}_p$.

There is a somewhat more direct way of introducing tensor fields. Let ξ^a be a contravariant vector field on M, and let f be a smooth function on M. Then, for each point p of M, ξ^a (p) is a contravariant vector at p. But, by definition (of a contravariant vector, Sect. 3) $\xi^a(p)$ assigns, to each smooth function on M, a real number. In particular, ξ^a (p) assigns a real number to our smooth function f. But this is true at each point p, and so we obtain a real number assigned to each point of M. In other words, we obtain a function on M. We write this function $\mathscr{Z}_\xi f$. Thus, given any contravariant vector field ξ^a on M and any smooth function f on M, $\mathscr{Z}_\xi f$ is a function on M. We can regard $\mathscr{Z}_\xi$ as a mapping from smooth function on M to functions on M. The conditions DD1, DD2, and DD3 (Sect. 3) become, respectively,

VF1. $\mathscr{Z}_\xi(f+g) = \mathscr{Z}_\xi f + \mathscr{Z}_\xi g$.
VF2. $\mathscr{Z}_\xi(fg) = f\mathscr{Z}_\xi g + g\mathscr{Z}_\xi f$.
VF3. If f = const. , then $\mathscr{Z}_\xi f = 0$.

Thus, we could just as well have defined a vector field as a mapping, $\mathscr{Z}_\xi$, from smooth functions on M to functions on M satisfying VF1, VF2, and VF3.

Next, let μ_a be a covariant vector field on M. Then, if ξ^a is any contravariant vector field on M, $\mu_a\xi^a$ is a function on M. Thus, the covariant vector field μ_a defines a mapping, $F(\xi^a)$, from contravariant vector fields on M to functions on M. If ξ^a and η^a are contravariant vector fields on M, and f is any smooth function on M, then, evidently

$$F(\xi^a + f\eta^a) = F(\xi^a) + fF(\eta^a) \tag{8}$$

Now let F be any mapping from contravariant vector fields on M to functions on M, and suppose that F is linear in the sense of (8). Then, at each point p of M, F defines a mapping from contravariant vectors at p to numbers, and, by (8), that mapping is linear. In other words, F assigns, to each point p of M, F defines a mapping from contravariant vectors at p to numbers, and, by (8), that mapping is linear. In other words, F assigns, to each point p of M, a covariant vector at p. That is, F defines a covariant vector field on M. We have shown that there is a one-to-one correspondence between covariant vector fields on M and linear mapping (linear in the sense of (8)) from contravariant vector fields on M to functions on M. Thus, we could just as well have defined a covariant vector field as a linear mapping (in the sense of (8)) from contravariant vector fields on M to functions on M.

Finally, let $\alpha^a{}_{bc}$ be a tensor field on M. Then, if μ_a, ξ^b, τ^c are vector fields on M, $\alpha^a{}_{bc}\,\mu_a\,\xi^b\,\tau^c$ is a function on M. So, $\alpha^a{}_{bc}$ defines a mapping $F(\mu_a, \xi^b, \tau^c)$, from certain

vector fields on M to functions on M. This mapping is multilinear, in the sense that, if f is any smooth function on M, then

$$\begin{aligned} F(\mu_a + f\nu_a, \xi^b, \tau^c) &= F(\mu_a, \xi^b, \tau^c) + fF(\nu_a, \xi^b, \tau^c) \\ F(\mu_a, \xi^b + f\eta^b, \tau^c) &= F(\mu_a, \xi^b, \tau^c) + fF(\mu_a, \eta^b, \tau^c) \\ F(\mu_a, \xi^b, \tau^c + f\sigma^c) &= F(\mu_a, \xi^b, \tau^c) + fF(\mu_a, \eta^b, \sigma^c) \end{aligned} \tag{9}$$

Conversely, a mapping $F(\mu_a, \xi^b, \tau^c)$ from certain vector fields on M to functions on M, multilinear in the sense of (9), defines a tensor field on M. Thus, we could just as well have defined the tensor fields on M as mapping, multilinear in the sense of (9), from vector fields on M to functions on M.

The purpose of the discussion above is to emphasize the following point: one could easily have defined the tensor fields directly without first defining tensors at a point. Contravariant vector fields can be defined in terms of their action (subject to VF1, VF2, and VF3) on smooth scalar fields; covariant vector fields in terms in their action (subject to (8)) on contravariant vector fields; tensor fields in terms of their action (subject to (9)) on vector fields. The discussions of Sections 3 and 5 can be regarded as motivation.

A contravariant vector field ξ^a is said to be *smooth* if, for every smooth function f, $\mathscr{L}_\xi f$ is a smooth function. A covariant vector field μ_a is said to be *smooth* if, for every smooth contravariant vector field ξ^a,$\mu_a\xi^a$ is a smooth function. A tensor field, e.g., $\alpha^a{}_{bc}$ is said to be *smooth* if, for any smooth vector fields μ_a, ξ^b, τ^c, $\alpha^a{}_{bc}\,\mu_a\tau^c$ is a smooth function. The smooth fields are the interesting ones. Hence, we shall hereafter normally omit the word "smooth": tensor fields are understood to be smooth unless otherwise stated. Smooth functions are sometimes called (smooth) *scalar fields*. It is clear from these definitions that the tensor operations – outer product, addition, index substitution, and contraction – produce smooth tensor fields when applied to smooth tensor fields.

Example 18. Let ξ^a be a (not necessarily smooth) contravariant vector field on M. Let (U, φ) be a chart. Let f be a smooth function on M, and set $\tilde{f} = f \cdot \varphi^{-1}, g = \mathscr{L}_\xi f$, and $\mathscr{L} = g \cdot \varphi^{-1}$. Let $\xi^1(x), \ldots, \xi^n(x)$ be the components of ξ^a with respect to this chart (Example 16). Then, from Eqn.5,

$$\tilde{g}(x) = \xi^1(x)\frac{\partial \tilde{f}}{\partial x^1} + \cdots + \xi^n(x)\frac{\partial \tilde{f}}{\partial x^n}$$

It follows immediately from this equation that a contravariant vector field is smooth if and only if its components with respect to every chart are C^∞ functions of $(x^1, \ldots, x^n)$.

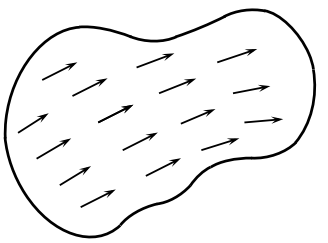
Exercise 31. Prove that a covariant vector field is smooth if and only if its components with respect to every chart are C^∞ functions. Same for an arbitrary tensor field.

The pattern above, an important one, will be repeated on several later occasions. One starts with something about the scalar fields, then extends it to the contravariant vector fields using the action $\mathscr{L}_\xi f$, then extends it to the covariant vector fields using the action $\xi^a\mu_a$, and finally extends it to all the tensor fields using the action $\alpha^a{}_{bc}\mu_a\xi^b\tau^c$.

7. Lie Derivatives

Let ξ^a be a (smooth) contravariant vector field, f a (smooth) scalar field, and consider the scalar field $\mathscr{Z}_\xi f$. Intuitively, ξ^a represents a "magnitude and direction" at each point of the manifold M. If, for each point of M, we take the "derivative of f" in the "direction of ξ^a", we obtain a a number. Doing so at each point, we obtain a scalar field on M, namely $\mathscr{Z}_\xi f$. So, $\mathscr{Z}_\xi f$ represents a sort of "directional derivative of f, at each point, in the direction of ξ^a". This $\mathscr{Z}_\xi f$ is sometimes called the *Lie derivative* of f in the ξ^a-direction. Is it possible to define a "directional derivative in the ξ^a-direction" for other things, e.g., for tensor (in addition to scalar) fields? The answer is yes. We do so in this section.

Let ξ^a and η^a be vector fields, f a scalar field. Then $\mathscr{Z}_\eta f$ is also a scalar field, so $\mathscr{Z}_\xi \mathscr{Z}_\eta f$ is still another scalar field. Set



$$\kappa(f) = \mathscr{Z}_\xi \mathscr{Z}_\eta f - \mathscr{Z}_\eta \mathscr{Z}_\xi f \tag{10}$$

so κ is a mapping from scalar fields (i.e., from $\mathfrak{F}$) to scalar fields. It follows immediately from VF1 that $\kappa(f+g) = \kappa(f) + \kappa(g)$. Furthermore, from VF1 and VF2,

$$\begin{aligned}
\kappa(fg) &= \mathscr{Z}_\xi \mathscr{Z}_\eta(fg) - \mathscr{Z}_\eta \mathscr{Z}_\xi(fg) \\
&= \mathscr{Z}_\xi(f\mathscr{Z}_\eta g + g\mathscr{Z}_\eta f) - \mathscr{Z}_\eta(f\mathscr{Z}_\xi g + g\mathscr{Z}_\xi f) \\
&= (\mathscr{Z}_\xi f)(\mathscr{Z}_\eta g) + f\mathscr{Z}_\xi\mathscr{Z}_\eta g + (\mathscr{Z}_\xi g)(\mathscr{Z}_\eta f) + g\mathscr{Z}_\xi\mathscr{Z}_\eta f \\
&\quad - (\mathscr{Z}_\eta f)(\mathscr{Z}_\xi g) - f\mathscr{Z}_\eta\mathscr{Z}_\xi g - (\mathscr{Z}_\eta g)(\mathscr{Z}_\xi f) - g\mathscr{Z}_\eta\mathscr{Z}_\xi f \\
&= f\,k(g) + g\,k(f)
\end{aligned}$$

Finally, from VP3, $\kappa(f) = 0$ if $f = \text{const}$. We have just shown that $\kappa(f)$ itself satisfies the three conditions VF1, VF2, and VF3. Hence, by the remarks of Sect. 6 (on defining contravariant vector fields), we have $\kappa(f) = \mathscr{Z}_\tau f$ (for all f) for some vector field τ^a. (Exercise 32. How do we know that τ^a is smooth?)

By definition, the *Lie derivative* of η^a in the ξ^a-direction, $\mathscr{Z}_\xi\eta^a$, is the vector field τ^a. In other words, we define $\mathscr{Z}_\xi\eta^a$ by the equation

$$\mathscr{Z}_\xi\mathscr{Z}_\eta f - \mathscr{Z}_\eta\mathscr{Z}_\xi f = \mathscr{Z}_{(\mathscr{Z}_\xi\eta)} f \tag{11}$$

for all f. The Lie derivative of a scalar field is a scalar field; the Lie derivative of a contravariant vector field is a contravariant vector field.

We have now extended the notion of the Lie derivative (in the ξ^a-direction) from scalar fields to contravariant vector fields. Let's work out some properties of it. First note that, for any scalar fields α and f, and vector fields ξ^a and λ^a,

$$\mathscr{Z}_{(\alpha\xi+\lambda)} = \alpha\mathscr{Z}_\xi f + \mathscr{Z}_\lambda f \tag{12}$$

Applying (12), in the case $\alpha = 1$, to the definition (11), it follows immediately that

$$\mathscr{Z}_\xi(\eta^a + \omega^a) = \mathscr{Z}_\xi\eta^a + \mathscr{Z}_\xi\omega^a \tag{13}$$

Now let α be a scalar field on M. What does $\mathscr{Z}_\xi(\alpha\eta^a)$ look like? It's easy to find out from the definition (11):

$$\begin{aligned}
&\mathscr{Z}_{(\mathscr{Z}_\xi\alpha\eta)} f = \mathscr{Z}_\xi\mathscr{Z}_{\alpha\eta} f - \mathscr{Z}_{\alpha\eta}\mathscr{Z}_\xi f \\
&= \mathscr{Z}_\xi(\alpha\mathscr{Z}_\eta f) - \alpha\mathscr{Z}_\eta\mathscr{Z}_\xi f \\
&= (\mathscr{Z}_\xi\alpha)(\mathscr{Z}_\eta f) + \alpha\mathscr{Z}_\xi\mathscr{Z}_\eta f - \alpha\mathscr{Z}_\eta\mathscr{Z}_\xi f \\
&= \mathscr{Z}_{(\eta\mathscr{Z}_\xi\alpha)} f + \alpha\mathscr{Z}_{(\mathscr{Z}_\xi\eta)} f \\
&= \mathscr{Z}_{(\eta\mathscr{Z}_\xi\alpha + \alpha\mathscr{Z}_\xi\eta)} f
\end{aligned} \tag{14}$$

Hence, since f is arbitrary,

$$\mathscr{Z}_\xi(\alpha\eta^a) = \eta^a \mathscr{Z}_\xi\alpha + \alpha\mathscr{Z}_\xi\eta^a \tag{15}$$

(Compare Eqn. (13) with VF1: Eqn. (15) with VP2.)

The next step is to define Lie derivatives of covariant vector fields. Let μ^a be a covariant vector field, and consider the scalar field

$$\mathscr{Z}_\xi(\eta^a\mu_a) - \mu_a\mathscr{Z}_\xi\eta^a \tag{16}$$

where η^a is any contravariant vector field. Replacing η^a in (16) by $\eta^a + \alpha\omega^a$ (α a scalar field),

$$\begin{aligned}
&\mathscr{Z}_\xi[(\eta^a + \alpha\omega^a)\mu_a] - \mu_a\mathscr{Z}_\xi(\eta^a + \alpha\omega^a)\\
&= \mathscr{Z}_\xi(\eta^a\mu_a) + \mathscr{Z}_\xi[\alpha(\omega^a\mu_a)] - \mu_a\mathscr{Z}_\xi\eta^a - \mu_a\mathscr{Z}_\xi(\alpha\omega^a)\\
&= \mathscr{Z}_\xi(\eta^a\mu_a) + \alpha\mathscr{Z}_\xi(\omega^a\mu_a) + (\omega^a\mu_a)\mathscr{Z}_\xi\alpha - \mu_a\mathscr{Z}_\xi\eta^a\\
&- \alpha\mu_a\mathscr{Z}_\xi\omega^a - (\mu_a\omega^a)\mathscr{Z}_\xi\alpha\\
&= [\mathscr{Z}_\xi(\eta^a\mu_a) - \mu_a\mathscr{Z}_\xi\eta^a] + \alpha[\mathscr{Z}_\xi(\omega^a\mu_a) - \mu_a\mathscr{Z}_\xi\omega^a]
\end{aligned}$$

where we have used (13), (15) and VF2. In other words, the expression (16) defines a linear (in the sense of (8)) mapping from contravariant vector fields (η^a) to scalar fields on M. By the remarks of Sect. 6 (on the definition of a covariant vector field), such a mapping represents a covariant vector field. We define the *Lie derivative* of μ_a (in the ξ^a−direction), $\mathscr{Z}_\xi\mu_a$, to be this vector field. That is, we define $\mathscr{Z}_\xi\mu_a$ by the equation

$$\eta^a\mathscr{Z}_\xi\mu_a = \mathscr{Z}_\xi(\eta^a\mu_a) - \mu_a\mathscr{Z}_\xi\eta^a \tag{17}$$

for all η^a.

We have now extended the notion of Lie derivative to covariant fields (the result again being a covariant vector field). Again, we want to work out some properties. It follows immediately from VF1, Eqn. (13), and the definition (17) that

$$\mathscr{Z}_\xi(\mu_a + \nu_a) = \mathscr{Z}_\xi\mu_a + \mathscr{Z}_\xi\nu_a \tag{18}$$

To evaluate $\mathscr{Z}_\xi(\alpha\mu_a)$, we use VF2 and the definition (17):

$$\begin{aligned}
\eta^a\mathscr{Z}_\xi(\alpha\mu_a) &= \mathscr{Z}_\xi(\alpha\mu_a\eta^a) - \alpha\mu_a\mathscr{Z}_\xi\eta^a\\
&= \mu_a\eta^a\mathscr{Z}_\xi\alpha + \alpha\mathscr{Z}_\xi(\mu_a\eta^a) - \alpha\mu_a\mathscr{Z}_\xi\eta^a\\
&= \eta^a[\mu_a\mathscr{Z}_\xi\alpha + \alpha\mathscr{Z}_\xi\mu_a]
\end{aligned}$$

Since this equation must hold for all η^a,

$$\mathscr{Z}_\xi(\alpha\mu_a) = \mu_a\mathscr{Z}_\xi\alpha + \alpha\mathscr{Z}_\xi\mu_a \tag{19}$$

The last step is to extend the notion of the Lie derivative from scalar and vector fields to arbitrary tensor fields. The method is the same. We merely do an example. Let $\alpha^a{}_{bc}$ be a tensor field, and consider the expression

$$\mathscr{Z}_\xi(\alpha^a{}_{bc}\mu_a\eta^b\lambda^c) - \alpha^a{}_{bc}\eta^b\lambda^c\mathscr{Z}_\xi\mu_a - \alpha^a{}_{bc}\mu_a\lambda^c\mathscr{Z}_\xi\eta^b - \alpha^a{}_{bc}\mu_a\eta^b\mathscr{Z}_\xi\lambda^c \tag{20}$$

Repeating (essentially unchanged) the calculation at the top of the preceding page, we see that the expression (20) is multilinear in μ_a, η^b, λ^c (in the sense of (9)). Hence, by the discussion of Sect. 6 (concerning definitions of tensor fields) the expression (20) defines a tensor field on M, which we denote by $\mathscr{Z}_\xi\alpha^a{}_{bc}$. That is to say, we define $\mathscr{Z}_\xi\alpha^a{}_{bc}$ by

$$\begin{aligned}
\mu_a\eta^b\lambda^c\mathscr{Z}_\xi\alpha^a{}_{bc} &= \mathscr{Z}_\xi(\alpha^a{}_{bc}\mu_a\eta^b\lambda^c)\\
&- \alpha^a{}_{bc}\eta^b\lambda^c\mathscr{Z}_\xi\mu_a - \alpha^a{}_{bc}\mu_a\lambda^c\mathscr{Z}_\xi\eta^b - \alpha^a{}_{bc}\mu_a\eta^b\mathscr{Z}_\xi\lambda^c
\end{aligned} \tag{21}$$

This is the *Lie derivative* of an arbitrary tensor field (in the ξ^a-direction). Linearity of the Lie derivative, i.e., the fact that

$$\mathscr{Z}_\xi(\alpha^a{}_{bc} + \beta^a{}_{bc}) = \mathscr{Z}_\xi\alpha^a{}_{bc} + \mathscr{Z}_\xi\beta^a{}_{bc} \tag{22}$$

follows immediately from VF1.

Eqns. (15) and (19) are suggestive. They can generalized. Let $\alpha^a{}_{bc} = \beta^a{}_b\gamma_c$ in (21). Then

$$\begin{aligned}
&\mu_a\eta^b\lambda^c\,\mathscr{Z}_\xi(\beta^a{}_b\gamma_c) = \mathscr{Z}_\xi(\beta^a{}_b\gamma_c\mu_a\eta^b\lambda^c)\\
&-\beta^a{}_b\gamma_c\eta^b\lambda^c\,\mathscr{Z}_\xi\mu_a - \beta^a{}_b\gamma_c\mu_a\lambda^c\,\mathscr{Z}_\xi\eta^b - \beta^a{}_b\gamma_c\mu_a\eta^b\,\mathscr{Z}_\xi\lambda^c\\
&= \beta^a{}_b\mu_a\eta^b\,\mathscr{Z}_\xi(\gamma_c\lambda^c) + \gamma_c\lambda^c\,\mathscr{Z}_\xi(\beta^a{}_b\mu_a\eta^b)\\
&-\beta^a{}_b\gamma_c\eta^b\lambda^c\,\mathscr{Z}_\xi\mu_a - \beta^a{}_b\gamma_c\eta_a\lambda^c\,\mathscr{Z}_\xi\eta^b - \beta^a{}_b\gamma_c\mu_a\eta^b\,\mathscr{Z}_\xi\lambda^c\\
&= (\beta^a{}_b\mu_a\eta^b)(\lambda^c\,\mathscr{Z}_\xi\gamma_c) + (\gamma_c\lambda^c)(\mu_a\eta^b\,\mathscr{Z}_\xi\beta^a{}_b)
\end{aligned}$$

Since μ_a,η^b, and λ^c are arbitrary, we have

$$\mathscr{Z}_\xi\beta^a{}_b\gamma_c = \beta^a{}_b\mathscr{Z}_\xi\gamma^c + \gamma_c\mathscr{Z}_\xi\beta^a{}_b \tag{23}$$

A similar equation holds, of course, for arbitrary outer products. Note that Eqns. (15) and (19) (as well as VF2) are special cases in which one factor in the outer products is a scalar field. Finally, rewriting Eqn. (17) in the form $\mathscr{Z}_\xi(\mu_a\eta^a) = \mu_a\mathscr{Z}_\xi\eta^a + \eta^a\mathscr{Z}_\xi\mu_a$, and noting the definition of contraction, we see that the contraction of the Lie derivative of any tensor field is equal to the Lie derivative of its contraction.

We summarize the properties of the Lie derivative in the following:

Theorem 8. Let ξ^a be a contravariant vector field on the manifold M. Then with any tensor field, $\alpha^{a\ldots c}{}_{b\ldots d}$, on M, there is associated a second tensor field, $\mathscr{Z}_\xi\alpha^{a\ldots c}{}_{b\ldots d}$, where this operation on tensor fields has the following properties:

LD1. $\mathscr{Z}_\xi(\alpha^{a\ldots c}{}_{b\ldots d} + \beta^{a\ldots c}{}_{b\ldots d} = \mathscr{Z}_\xi(\alpha^{a\ldots c}{}_{b\ldots d}) + \mathscr{Z}_\xi(\beta^{a\ldots c}{}_{b\ldots d})$.

LD2. $\mathscr{Z}_\xi(\alpha^{a\ldots c}{}_{b\ldots d}\gamma^{m\ldots n}{}_{p\ldots q}) = \alpha^{a\ldots c}{}_{b\ldots d}\mathscr{Z}_\xi(\gamma^{m\ldots n}{}_{p\ldots q})$
$+\gamma^{m\ldots n}{}_{p\ldots q}\mathscr{Z}_\xi(\alpha^{a\ldots c}{}_{b\ldots d})$.

LD3. The operations $\mathscr{Z}_\xi$ and index substitution can be performed in either order, with same result.

LD4. The operations $\mathscr{Z}_\xi$ and construction can be performed in either order, with the same result.

Property LD2 is sometimes called the Leibnitz rule. That there are four properties in Theorem 4 is no coincidence: there are four fundamental operations on tensors (addition, outer product, index substitution, and contraction), so the properties above express the interaction between Lie derivatives and these tensor operations. Note that practically every interesting equation in this section is a consequence of the four properties above, e.g., VF1, VF2, (13), (15), (17), (18), (19), (21), and (23). Although the definition of the Lie derivative is, of course, worth knowing, it is the properties above which are most useful.

Two further remarks about the Lie derivative should be noted. Since Lie derivatives are always in the direction of a contravariant vector field, the Lie derivatives of contravariant vector fields have an especially pretty structure. Let ξ^a and η^a be contravariant vector fields, and, for the purposes of this paragraph, set $[\xi,\eta] = \mathscr{Z}_\xi\eta^a$. Then, from (11), we have immediately

$$[\xi,\eta] = -[\eta,\xi] \tag{24}$$

Furthermore, we have, for three contravariant vector fields,

$$[\lambda,[\xi,\eta]] + [\xi,[\eta,\lambda]] + [\eta,[\lambda,\xi]] = 0 \tag{25}$$

(Proof: Apply the left side of (25) to an arbitrary scalar field f. Now expand, using (11), and note that all the terms cancel.) So, on the contravariant vector fields on a manifold, there is defined a skew (Eqn. (24)) "bracket" which obeys Jacobi's identity (Eqn. (25)). That is to say, the contravariant vector fields on a manifold have the structure of a Lie algebra.

The second remark concerns the dependence of the Lie derivative on the vector field representing the direction in which the Lie derivative is taken. (Theorem 8 refers to the dependence on the tensor field which is "Lie derived".) We need a bit of notation. Fix a scalar field α. Then, foe each ξ^a, $\mathscr{Z}_\xi\alpha$ is a scalar field. By (12), this scalar field is linear (in the sense of (8)) in ξ^a. Hence, α defines a covariant vector field. This vector field, which is called the *gradient* of α, is written $D(\alpha)_a$. Thus, $D(\alpha)_a$ is defined by

$$\xi^a D(\alpha)_a = \mathscr{Z}_\xi\alpha \tag{26}$$

For Lie derivatives of scalar fields, the dependence on ξ^a is given by (12). For contravariant vector fields, the dependence follows immediately from Eqn. (15) and the fact that $\mathscr{Z}_\xi\eta^a = \mathscr{Z}_\eta\xi^a$:

$$\mathscr{Z}_{(\alpha\xi+\lambda)}\eta^a = \alpha\mathscr{Z}_\xi\eta^a + \mathscr{Z}_\lambda\eta^a - \xi^a\eta^m D(\alpha)_m \tag{27}$$

For covariant vector fields, we use Eqn. (17):

$$\begin{aligned}\tau^a\mathscr{Z}_{(\alpha\xi+\lambda)}\mu_a &= \mathscr{Z}_{(\alpha\xi+\lambda)}(\tau^a\mu_a) - \mu_a\mathscr{Z}_{(\alpha\xi+\lambda)}\tau^a\\ &= \alpha\mathscr{Z}_\xi(\tau^a\mu_a) + \mathscr{Z}_\lambda(\tau^a\mu_a) - \mu_a\alpha\mathscr{Z}_\xi\tau^a - \mu_a\mathscr{Z}_\lambda\tau^a + \mu_a\xi^a\tau^b D(\alpha)_b\\ &= \tau^a\alpha\mathscr{Z}_\xi\mu_a + \tau^a\mathscr{Z}_\lambda\mu_a + \mu_a\xi^a\tau^b D(\alpha)_b\end{aligned}$$

Since this holds for all τ^a, we have

$$\mathscr{Z}_{(\alpha\xi+\lambda)}\mu_a = \alpha\mathscr{Z}_\xi\mu_a + \mathscr{Z}_\lambda\mu_a + D(\alpha)_a\mu_b\xi^b \tag{28}$$

Exercise 33. Prove that, for an arbitrary tensor field, e.g., $\beta^a{}_{bc}$, we have

$$\begin{aligned}\mathscr{Z}_{(\alpha\xi+\lambda)}\beta^a{}_{bc} = \alpha\mathscr{Z}_\xi\beta^a{}_{bc} + \mathscr{Z}_\lambda\beta^a{}_{bc} - \xi^a\beta^m{}_{bc}D(\alpha)_m\\ + D(\alpha)_b\beta^a{}_{mc}\xi^m + D(\alpha)_c\beta^a{}_{bm}\xi^m\end{aligned} \tag{29}$$

(Hint: Use (27), (28), and (21).) Note that (12), (27), and (28) are special cases of (29).

Exercise 34. Express Lie derivatives in terms of components.

Example 19. Let ξ^a and μ_a be vector fields. Then $\xi^a\mu_a$ is a scalar field. Hence, we have a multilinear mapping (in the sense of (9)) from smooth ξ^a, μ_a to smooth scalar fields. So we have defined a tensor field, which we write $\delta^a{}_b$ (See Example 15.) That is, $\delta^a{}_b$ is defined by $\delta^a{}_b\mu_a\xi^b = \mu_a\xi^a$, for every μ_a, ξ^b. (This $\delta^a{}_b$ is called the *unit tensor*.) We evaluate its Lie derivative. From Eqn. (21).

$$\begin{aligned}\alpha_a\beta^b\mathscr{Z}_\xi\delta^a{}_b &= \mathscr{Z}_\xi(\delta^a{}_b\alpha_a\beta^b) - \delta^a{}_b\beta^b\mathscr{Z}_\xi\alpha_a - \delta^a{}_b\alpha_a\mathscr{Z}_\xi\beta^b\\ &= \mathscr{Z}_\xi(\alpha_a\beta^a) - \beta^a\mathscr{Z}_\xi\alpha_a - \alpha\mathscr{Z}_\xi\beta^a\end{aligned}$$

But now, from Eqn, (17), we have $\mathscr{Z}_\xi\delta^a{}_b = 0$.

Exercise 35. Find the components of $\delta^a{}_b$ with respect to a chart, and verify the conclusion of Example 19 using your formula of Exercise 34.

Exercise 36. A constant scalar field has the property that its Lie derivative in the direction of every vector field is zero. Thus, one might be tempted to define a "constant tensor field" as one whose Lie derivative in the direction of every vector field is zero. Show that this idea does not work, in the following sense: a "constant tensor field" of rank one necessarily vanishes.

Example 20. Let ξ^a be a vector field, α a scalar field. We prove that

$$D(\mathscr{Z}_\xi \alpha)_a = \mathscr{Z}_\xi D(\alpha)_a \tag{30}$$

To do this note that

$$\begin{aligned}\eta^a \mathscr{Z}_\xi D(\alpha)_a &= \mathscr{Z}_\xi \eta^a D(\alpha)_a - D(\alpha)_a \mathscr{Z}_\xi \eta^a = \mathscr{Z}_\xi \mathscr{Z}_\eta \alpha - \mathscr{Z}_{(\mathscr{Z}_\xi \eta)} \alpha \\ &= \mathscr{Z}_\eta \mathscr{Z}_\xi \alpha = \eta^a D(\mathscr{Z}_\xi \alpha)_a\end{aligned}$$

where we have used (11). Since η^a is arbitrary, (30) follows. (That is, the Lie derivative commutes with the gradient.) Note that Eqn. (30) implies Eqn. (11).

Exercise 37. Can $\mathscr{Z}_\xi$ be defined in any reasonable way when ξ^a is not smooth?

Exercise 38. Find a list of axioms on an operator K on tensors which are necessary and sufficient for $K = \mathscr{Z}_\eta$ for some vector field ξ^a.

Exercise 38. Let ξ^a and η^a be contravariant vector fields. Prove that

$$\mathscr{Z}_\xi \mathscr{Z}_\eta \alpha^{a\ldots c}{}_{b\ldots d} - \mathscr{Z}_\eta \mathscr{Z}_\xi \alpha^{a\ldots c}{}_{b\ldots d} = \mathscr{Z}_{(\mathscr{Z}_\xi \eta)} \alpha^{a\ldots c}{}_{b\ldots d} \tag{31}$$

(This equation generalizes (11).) (For scalars, (31) follows from (11). For contravariant vectors, contract (31) with $D(f)_a$, f arbitrary, and verify using (11). Now extend to covariant vectors, and finally to tensors of arbitrary rank.)

We shall obtain a more geometrical interpretation of the Lie derivative later.

8. Symmetrization and Antisymmetrization

Let $\alpha_a{}^m{}_{bc}{}^d$ and $\beta^{mn}{}_a{}^{cb}$ be tensor fields. Set

$$\alpha_{(a}{}^m{}_{b)c}{}^d = \frac{1}{2}(\alpha_a{}^m{}_{bc}{}^d + \alpha_b{}^m{}_{ac}{}^d)$$
$$\alpha_{[a}{}^m{}_{b]c}{}^d = \frac{1}{2}(\alpha_a{}^m{}_{bc}{}^d - \alpha_b{}^m{}_{ac}{}^d)$$
$$\beta^{m(n}{}_a{}^{cb)} = \frac{1}{6}(\beta^{mn}{}_a{}^{cb} + \beta^{mc}{}_a{}^{bn} + \beta^{mb}{}_a{}^{nc} + \beta^{mn}{}_a{}^{bc} + \beta^{mc}{}_a{}^{nb} + \beta^{mb}{}_a{}^{cn})$$
$$\beta^{m[n}{}_a{}^{cb]} = \frac{1}{6}(\beta^{mn}{}_a{}^{cb} + \beta^{mc}{}_a{}^{bn} + \beta^{mb}{}_a{}^{nc} - \beta^{mn}{}_a{}^{bc} - \beta^{mc}{}_a{}^{nb} - \beta^{mb}{}_a{}^{cn})$$

More generally, a tensor, written with round brackets surrounding a collection of p consecutive indices (all covariant or all contravariant), means $1/p!$ times the sum of the $p!$ tensors obtained (using index substitution) by arranging those p indices in all possible orders. A tensor, written with square brackets surrounding a collection of p consecutive indices (all covariant or all contravariant), means $1/p!$ times a linear combination of the $p!$ tensors obtained by arranging those p indices in all possible orders, where a plus sign appears before those tensors in which the arrangement of indices is an even permutation of the original index arrangement, and a minus sign appears otherwise. These operations are called *symmetrization* and *antisymmetrization* (over the indices in question), respectively. These definitions occasionally make it possible to write expressions more concisely.

Example 21. $\tau_{[bac]} = -\tau_{[abc]} = -\tau_{[bca]} = \tau_{[cba]}$. $\tau_{(bac)} = \tau_{(abc)} = \tau_{(bca)} = \tau_{(cba)}$

Example 22. Set $\alpha_{abcdef} = \beta_{ab[cde]f}$. Then $\alpha_{a[bcdef]} = \beta_{a[bcdef]}$, while $\alpha_{a[bcdef]} = 0$

Example 23. $\xi^m\alpha_{[ma}\beta_{b]} = 2/3\xi^m\alpha_{m[a}\beta_{b]} + 1/3\alpha_{ab}\xi^m\beta_m$.

Example 24. If $\tau_{bac} = -\tau_{abc}$, $\tau_{acb} = -\tau_{abc}$, and $\tau_{cba} = -\tau_{abc}$, then $\tau_{abc} = \tau_{[abc]}$. More generally, $\alpha_{a...c} = \alpha_{[a...c]}$ if and only if reverses sign under interchange of any two of its indices.

Example 25. $\beta_{a...c} = \beta_{(a...c)}$ if and only if is invariant under interchange of any two of its indices.

Example 26. If $\xi^a\eta^b\alpha_{ab} = -\xi^b\eta^a\alpha_{ab}$ for all ξ^a, η^a then $(\alpha_{ab} + \alpha_{bc})\xi^a\eta^b = 0$ for all ξ^a, η^a, whence $(\alpha_{ab} + \alpha_{ab}) = 0$, whence $\alpha_{ab} = \alpha_{[}ab]$.

9. Exterior Derivatives

We now introduce a second type of "natural derivative" on manifolds. We summarize its structure:

	Lie Derivative	*Exterior Derivative*
Generalization of	Directional Derivative	Gradient
Requires	A Contravariant Vector Field	Nothing
Applicable to	All Tensor Fields	Tensor Fields $\omega_{a\ldots c} = \omega_{[a\ldots c]}$
Index Structure of Result	Same as Original Tensor Field	Adds One Index

Thus, the exterior derivative has the advantage over the Lie derivative that no extraneous tensor fields (i.e., ξ^a) are required, and the disadvantage that it is applicable only to tensor fields with a certain index structure. We shall see later that Lie derivatives and exterior derivatives are special cases (by far the most common and most useful ones) of a more general class of objects.

The *exterior derivative* of a scalar field, ω, is defined as its gradient, $D(\omega)_m$. (See Eqn. (30).) It follows immediately from VF2 that

$$D(\mu\omega)_m = \omega D(\mu)_m + \mu D(\omega)_m \tag{32}$$

where μ is any scalar field.

Now let ω_a be a contravariant vector field. Set $\lambda = \xi^a\omega_a$, and consider the right side of

$$\xi^m D(\omega)_{ma} = \frac{1}{2}\mathscr{L}_\xi\omega_a = \frac{1}{2}D(\lambda)_a \tag{33}$$

Replacing ξ^a by $\alpha\xi^a$ in the right side of (33), we have

$$\begin{aligned}\frac{1}{2}\mathscr{L}_{\alpha\xi}\omega_a - \frac{1}{2}D(x\lambda)_a &= \frac{1}{2}\alpha\mathscr{L}_\xi\omega_a + \frac{1}{2}D(\alpha)_a\xi^m\omega_m - \frac{1}{2}\alpha D(\lambda)_a - \frac{1}{2}\lambda D(\alpha)_a \\ &= \alpha\left[\frac{1}{2}\mathscr{L}_\xi\omega_a - D(\lambda)_a\right]\end{aligned}$$

were we have used (32) and (28). Hence, Eqn. (33) defines a tensor field, $D(\omega)_{ma}$. This $D(\omega)_{ma}$ is the *exterior derivative* of the covariant vector field ω_a. To show that $D(\omega)_{ma}$ is antisymmetric, contract (33) with an arbitrary η^a:

$$\begin{aligned}\eta^a\xi^m D(\omega)_{ma} &= \frac{1}{2}\eta^a\mathscr{L}_\xi\omega_a - \frac{1}{2}\eta^a D(\lambda)_a \\ &= \frac{1}{2}\mathscr{L}_\xi(\eta^a\omega_a) - \frac{1}{2}\omega_a\mathscr{L}_\xi\eta^a - \frac{1}{2}\mathscr{L}_\eta(\xi^a\omega_a)\end{aligned}$$

and note that the last expression above reverses sign under interchange of ξ^a and η^a. (The first and third terms together reverse sign, and the second term alone reverses sign.) Since ξ^a and η^a are arbitrary,

$$D(\omega)_{ma} = D(\omega)_{[ma]} \tag{34}$$

Finally, replacing ω_a by $\mu\omega_a$ in (33), we have

$$\begin{aligned}\xi^m D(\mu\omega)_{ma} &= \frac{1}{2}\mathscr{Z}_\xi(\mu\omega_a) - \frac{1}{2}D(\mu\lambda)_a \\ &= \frac{1}{2}\mu\mathscr{Z}_\xi\omega_a + \frac{1}{2}\omega_a\mathscr{Z}_\xi\mu - \frac{1}{2}\lambda D(\mu)_a - \frac{1}{2}\mu D(\lambda)_a \\ &= \mu\xi^m D(\omega)_{ma} + \frac{1}{2}\xi^m(D(\mu)_m\omega_a - \omega_m D(\mu)_a)\end{aligned}$$

Since ξ^m is arbitrary,

$$D(\mu\omega)_{ma} = \mu D(\mu)_{ma} + D(\mu)_{[m}\omega_{a]} \tag{35}$$

Now let $\omega_{ab} = \omega_{[ab]}$ be a tensor field. Set $\lambda_b = \xi^a\omega_{ab}$, and consider the right side of

$$\xi^m D(\omega)_{mab} = \frac{1}{3}\mathscr{Z}_\xi\omega_{ab} - \frac{2}{3}D(\lambda)_{ab} \tag{36}$$

Replacing ξ^m by $\alpha\xi^m$ in the right side of (36),

$$\begin{aligned}&\frac{1}{3}\mathscr{Z}_{\alpha\xi}\omega_{ab} - \frac{2}{3}D(\alpha\lambda)_{ab} = \frac{1}{3}\alpha\mathscr{Z}_\xi\omega_{ab} + \frac{1}{3}D(\alpha)_a\xi^m\omega_{mb} + \frac{1}{3}D(\alpha)_b\xi^m\omega_{am} \\ &- \frac{2}{3}\alpha D(\lambda)_{ab} - \frac{2}{3}D(\alpha)_{[a}\lambda_{b]} = \alpha\Big[\frac{1}{3}\mathscr{Z}_\xi\omega_{ab} - \frac{2}{3}D(\lambda)_{ab}\Big]\end{aligned}$$

Hence, Eqn. (36) defines a tensor field, $D(\omega)_{mab}$, which we call the *exterior derivative* of ω_{ab}. Contracting (36) with η^a,

$$\begin{aligned}\eta^a\xi^m D(\omega)_{mab} &= \frac{1}{3}\eta^a\mathscr{Z}_\xi\omega_{ab} - \frac{2}{3}\eta^a D(\lambda)_{ab} \\ &= \frac{1}{3}\mathscr{Z}_\xi\eta^a\omega_{ab} - \frac{1}{3}\omega_{ab}\mathscr{Z}_\xi\eta^a - \frac{2}{3}\Big(\frac{1}{2}\mathscr{Z}_\eta\lambda_b - \frac{1}{2}D(\kappa)_b\Big) \\ &= \frac{1}{3}\mathscr{Z}_\xi(\eta^a\omega_{ab}) - \frac{1}{3}\mathscr{Z}_\eta(\xi^a\omega_{ab}) - \frac{1}{3}\omega_{ab}\mathscr{Z}_\xi\eta^a + \frac{1}{3}D(\kappa)_b\end{aligned}$$

where we have set $\kappa = \xi^a\eta^b\omega_{ab}$. Reversing the roles of ξ^a and η^a, the first two terms in the last expression above reverse sign. Since, furthermore, κ reverses sign, the entire expression reverses sign. Hence, since ξ^a and η^a are arbitrary, $D(\omega)_{mab}$ is antisymmetric in m and a. Since, furthermore, $D(\omega)_{mab}$ is (clearly, from(36)) antisymmetric in a and b, we have

$$D(\omega)_{mab} = D(\omega)_{[mab]} \tag{37}$$

Finally, replacing ω_{ab} in (36) by $\mu\omega_{ab}$,

$$\begin{aligned}\xi^m D(\mu\omega)_{mab} &= \frac{1}{3}\mathscr{Z}_\xi(\mu\omega_{ab}) - \frac{2}{3}D(\mu\lambda)_{ab} \\ &= \mu\Big(\frac{1}{2}\mathscr{Z}_\xi\omega_{ab} - \frac{2}{3}D(\lambda)_{ab}\Big) + \frac{1}{3}\omega_{ab}\mathscr{Z}_\xi\mu - \frac{2}{3}D(\mu)_{[a}\lambda_{b]} \\ &= \mu D(\omega)_{mab}\xi^m + \xi^m D(\mu)_{[m}\omega_{ab]}\end{aligned}$$

Since ξ^m is arbitrary, we have

$$D(\mu\omega)_{mab} = \mu D(\omega)_{[m}\omega_{ab]} \tag{38}$$

We proceed by induction (on the rank of ω). Let $\omega_{a\ldots c} = \omega_{[a\ldots c]}$ have rank p. Define the *exterior derivative* of $\omega_{a\ldots c}$ by

$$\xi^m D(\omega)_{ma\ldots c} = \frac{1}{p+1}\mathscr{Z}_\xi\omega_{a\ldots c} - \frac{p}{p+1}D(\lambda)_{a\ldots c}\lambda_{b\ldots c} = \xi^a\omega_{ab\ldots c} \tag{39}$$

Then, as above, one shows that

$$D(\omega)_{ma\ldots c} = D(\omega)_{[ma\ldots c]} \tag{40}$$

$$D(\mu\omega)_{ma\ldots c} = D(\mu)_{[m}\omega_{a\ldots c]} \tag{41}$$

A tensor field with p covariant indices and no contravariant indices, which is antisymmetric in all its indices, is called a *p-form*. Thus, the exterior derivative of a p-form is a (p+1)-form.

We next establish two further properties of the exterior derivative. Let $\omega_{a\ldots c}$ be a p-form. Then

$$\mathscr{Z}_\xi D(\omega)_{ma\ldots c} = D(\mathscr{Z}_\xi\omega)_{ma\ldots c} \tag{42}$$

$$D(D(\omega))_{mna\ldots c} = 0 \tag{43}$$

where, in (42), ξ^a is arbitrary. Eqn. (42) states that exterior derivatives and Lie derivatives commute. Eqn. (43) states that, if exterior differentiation is applied twice in succession, the result is zero. (The exterior derivative is often called the curl.) To prove (42) and (43), we again proceed by induction on p. First note (Example 30) that (42) holds when ω is a scalar field. Using this fact, we have, for ω a scalar field,

$$\begin{aligned}\xi^m D(D(\omega))_{ma} &= \frac{1}{2}\mathscr{Z}_\xi D(\omega)_a - \frac{1}{2}D(\lambda)_a \\ &= \frac{1}{2}D(\mathscr{Z}_\xi\omega)_a - \frac{1}{2}D(\lambda)_a = 0\end{aligned}$$

where $\lambda = \xi^a D(\omega)_a = \mathscr{Z}_\xi\omega$. Hence, since ξ^m is arbitrary, (43) is satisfied for ω a scalar field. Now let ω_a be a vector field. Then

$$\begin{aligned}&\eta^m \mathscr{Z}_\xi D(\omega)_{ma} - \eta^m D(\mathscr{Z}_\xi\omega)_{ma} \\ &= \mathscr{Z}_\xi(\eta^m D(\omega)_{ma}) - D(\omega)_{ma}\mathscr{Z}_\xi\eta^m - \frac{1}{2}\mathscr{Z}_\eta\mathscr{Z}_\xi\omega_a + \frac{1}{2}D(\lambda)_a \\ &= \mathscr{Z}_\xi\Big(\frac{1}{2}\mathscr{Z}_\eta\omega_a - \frac{1}{2}D(\mu)_a\Big) - \Big(\frac{1}{2}\mathscr{Z}_{(\mathscr{Z}_\xi\eta)}\omega_a - \frac{1}{2}D(\nu)_a\Big) \\ &- \frac{1}{2}\mathscr{Z}_\eta\mathscr{Z}_\xi\omega_a + \frac{1}{2}D(\lambda)_a \\ &= -\frac{1}{2}\mathscr{Z}_\xi D(\mu)_a + \frac{1}{2}D(\nu)_a + \frac{1}{2}D(\lambda)_a \\ &= \frac{1}{2}D(-\mathscr{Z}_\xi\mu + \nu + \lambda)_a = 0\end{aligned}$$

where we have set $\lambda = \eta^m\mathscr{Z}_\xi\omega_m, \mu = \eta^a\omega_a$, and $\nu = \omega_a\mathscr{Z}_\xi\eta^a$. Since η^a is arbitrary, (42) is satisfied for a 1-form. Therefore, for ω_a,

$$\begin{aligned}\xi^m D(D(\omega))_{mna} &= \frac{1}{3}\mathscr{Z}_\xi D(\omega)_{na} - \frac{2}{3}D(\lambda)_{na} \\ &= \frac{1}{3}D(\mathscr{Z}_\xi\omega)_{na} - \frac{2}{3}D(\lambda)_{na} = \frac{1}{3}D(\kappa)_{na}\end{aligned}$$

where $\lambda_a = \xi^m D(\omega)_{ma}$, and

$$\kappa_a = \mathscr{Z}_\xi\omega_a - 2\xi^m D(\omega)_{ma} = D(\xi^m\omega_m)_a$$

Therefore, (43) is satisfied for ω_a a 1-form. Continuing in this way, proving first (42) and then (43) for each rank, we see that (42) and (43) are true in general.

One further property of the exterior derivative remains to be shown. Eqn. (38) suggests that the exterior derivative satisfies a "Leibnitz rule" under outer product. However,

the outer product of two forms, e.g., $\alpha_{a\ldots c}\beta_{d\ldots f}$, is not in general antisymmetric. (In fact, it is antisymmetric if and only if one factor is 0−form or one factor is zero.) To correct this, we take the outer product and antisymmetrize, i.e., $\alpha_{[a\ldots c}\beta_{d\ldots f]}$ This is called the wedge product of α and β. We now prove the "generalized Leibnitz rule" for exterior derivatives:

$$D(\omega)_{ma\ldots f} = D(\alpha)_{[ma\ldots c}\beta_{d\ldots f]} + D(\beta)_{[md\ldots f}\alpha_{a\ldots c]} \tag{44}$$

where $\omega_{a\ldots f} = \alpha_{[a\ldots c}\beta_{d\ldots f]}$. Let p and q be the ranks of $\alpha_{a\ldots c}$ and $\beta_{d\ldots f}$, respectively, and set

$$\mu_{b\ldots f} = \xi^a\alpha_{ab\ldots c} \quad \nu_{c\ldots f} = \xi^d\beta_{dc\ldots f}$$

$$\lambda_{b\ldots f} = \xi^a\alpha_{a\ldots f} = \frac{p}{p+q}\mu_{[b\ldots c}\beta_{d\ldots f]} + \frac{(-1)^{p-1}q}{p+q}\alpha_{[b\ldots cd}\nu_{e\ldots f]}$$

Then

$$\begin{aligned}
&\xi^m D(\omega)_{ma\ldots f} = \frac{1}{p+q+1}\mathscr{Z}_\xi\omega_{a\ldots f} - \frac{p+q}{p+q+1}(D)\lambda_{ab\ldots f}\\
&= \frac{1}{p+q+1}\alpha_{[a\ldots c}\mathscr{Z}_\xi\beta_{d\ldots f]} + \frac{1}{p+q+1}\beta_{[d\ldots f}\mathscr{Z}_\xi\alpha_{a\ldots c]} - \frac{p}{p+q+1}D(\nu)_{[a\ldots c}\beta_{d\ldots f]}\\
&- \frac{p}{p+q+1}D(\beta)_{[ad\ldots f}\mu_{b\ldots c]} - \frac{(-1)^{p-1}q}{p+q+1}D(\alpha)_{[ab\ldots d}\nu_{e\ldots f]} - \frac{(-1)^{p-1}q}{p+q+1}D(\nu)_{[ac\ldots f}\alpha_{b\ldots d]}\\
&= \xi^m\Big(D(\alpha)_{[ma\ldots c}\beta_{d\ldots f]} + D(\beta)_{[md\ldots f}\alpha_{a\ldots c]}\Big)
\end{aligned}$$

which proves (44)

We summarize:

Theorem 9. The exterior derivative satisfies

ED1. $D(\omega+\mu)_{ma\ldots d} = D(\omega)_{ma\ldots d} + D(\mu)_{ma\ldots d}$

ED2. $D(\omega)_{ma\ldots d} = D(\alpha)_{[ma\ldots b}\beta_{c\ldots d]} + D(\beta)_{[mc\ldots d}\alpha_{a\ldots b]}(\omega_{a\ldots d} = \alpha_{[a\ldots b}\beta_{c\ldots d]})$

ED3. $D(D(\omega))_{mna\ldots d} = 0$

Furthermore, the exterior derivative and Lie derivative are related by

ELD1. $\mathscr{Z}_\xi D(\omega)_{ma\ldots d} = D(\mathscr{Z}_\xi\omega)_{ma\ldots d}$

ELD2. $\mathscr{Z}_\xi\omega_{a\ldots d} - pD(\lambda)_{a\ldots d} - (p+1)\xi^m D(\omega)_{ma\ldots d} = 0$

where, in ELD2, we have set $\lambda_{b\ldots d} = \xi^a\omega_{ab\ldots d}$, and where $\omega_{a\ldots d}$ is a p−form.

Exercise 39. Write the exterior derivative in terms of components.

Example 27. In electrodynamics, the vector potential is a 1-form, A_b. The electromagnetic field is given by $F_{ab} = D(A)_{ab}$. It follows from ED3 that $D(F)_{mab} = 0$. This is one of Maxwell's equations.

Example 28. In classical mechanics, the symplectic structure on phase space is given by a 2-form, S_{ab}, with $D(S)_{mab} = 0$. A vector field ξ^a on phase space space is said to generate an infinitesimal canonical transformation if $\mathscr{Z}_\xi S_{ab} = 0$. Setting $\lambda_b = \xi^a S_{ab}$, we see from ELD2 that, if ξ^a generates an infinitesimal canonical transformation, then $D(\lambda)_{mb} = 0$. The Hamiltonian H is a function on phase space. Choose ξ^a such that $\xi^a S_{ab} = D(H)_b$. Then, from ED3, ξ^a automatically generates an infinitesimal canonical transformation. This transformation describes the dynamics of the system.

10. Derivative Operators

We have now seen two contexts in which derivatives are applicable to tensor fields: Lie derivatives and exterior derivatives. The first of these requires a vector field ξ^a; the second is applicable only to tensor fields with a certain index structure. The time has now come to ask about derivative-type operations which satisfy all the nice properties one would ask of such an operation.

By a *derivative operator*, we understand a mapping, ∇_a, from tensor fields to tensor fields, where $\nabla_a \alpha_{m\ldots}{}^{c}$ has one more covariant index than $\alpha_{m\ldots}{}^{c}$, this mapping subject to the following conditions:

DO1. $\nabla_a(\alpha_{m\ldots}{}^{c} + \beta_{m\ldots}{}^{c}) = \nabla_a \alpha_{m\ldots}{}^{c} + \nabla_a \beta_{m\ldots}{}^{c}$.

DO2. $\nabla_a(\alpha_{m\ldots}{}^{c} \mu^{n\ldots}{}_{d}) = \alpha_{m\ldots}{}^{c} \nabla_a \mu^{n\ldots}{}_{d} + \mu^{n\ldots}{}_{d} \nabla_a \alpha_{m\ldots}{}^{c}$.

DO3. The derivative operator ∇_a can applied before or after contraction, with the same result. The derivative operator ∇_a can be applied before or after index substitution, with the same result.

DO4. For any scalar field α, $\nabla_a \alpha = D(\alpha)_a$.

DO5. For any scalar field α, $\nabla_{[a} \nabla_{b]} \alpha = 0$.

Of course, derivative operators ∇_b, ∇_c, etc. are induced from ∇_a by index substitution. The operator ∇_a cannot be applied to a tensor field having "a" as a covariant index, i.e., we cannot write $\nabla_a \alpha_m{}^{bc}{}_{ap}$. If ∇_a is applied to a tensor field with "a" as a contravariant index, then contraction is understood. That is, if $\beta_{am}{}^{bc}{}_{p} = \nabla_a \alpha_m{}^{bc}{}_{p}$. Then we write $\nabla_a \alpha_m{}^{ba}{}_{p}$ for $\beta_{am}{}^{ba}{}_{p}$. Condition DO1 is, of course, natural for a derivative operator, while DO2 is the Leibnitz rule again. Conditions DO1 and DO2 are characteristic of things which "behave like derivatives". In more detail, the first statement in condition DO3 means the following. If $\beta_{am}{}^{bc}{}_{p} = \nabla_a \alpha_m{}^{bc}{}_{p}$, then $\beta_{am}{}^{bm}{}_{p} = \nabla_a(\alpha_m{}^{bm}{}_{p})$. There is only one reasonable type of "derivative" of a scalar field, namely, the gradient. Condition DO4 ensures that the derivative operator, applied to a scalar field, is just this gradient. Condition DO5 states that two derivative operators can be applied to a scalar field in either order (i.e., $\nabla_a \nabla_b \alpha = \nabla_b \nabla_a \alpha$). That is to say, derivatives commute on scalar fields. One could require DO5 for all tensor fields, but this turns out to eliminate most interesting and useful cases. Occasionally, one sees the notion of a derivative operator introduced with condition DO5 omitted. These "generalized derivative operators" are said to have torsion, and, when DO5 happened to be satisfied, the operator is said to be torsion-free. We shall see shortly that "derivative operators with torsion" are easily treated within the context of the definition above.

Example 29. Consider the $n-$dimensional manifold R^n, with its natural chart (Example 1). Then a tensor field is uniquely characterized by its components with respect to this chart. If, for example, $\alpha_a{}^{bc}$ is a tensor field, and $\alpha_i{}^{jk}(x)$ $(i, j, k = 1, 2, \ldots, n)$ are its components, then denote by $\nabla_d \alpha_a{}^{bc}$ the tensor field whose components, with respect to this chart, are $\frac{\partial}{\partial x^m} \alpha_i{}^{jk}(x)$ $(i, j, k, m = 1, 2, \ldots, n)$. This ∇_d is a derivative operator on R^n, i.e., it satisfies DO1–DO5.

The natural question to ask, in response to the definition above, are those concerning the existence and uniqueness of derivative operators on a given manifold M. These questions have complete, and relatively simple, answers. We begin with uniqueness.

Let ∇_a and ∇'_a be two derivative operators. Then, for any covariant vector fields μ_b and ν_b, and any scalar field α, we have

$$\begin{aligned}(\nabla'_a - \nabla_a)(\alpha\mu_b + \nu_b) &= \mu_b(\nabla'_a - \nabla_a)\alpha + \alpha(\nabla'_a - \nabla_a)\nu_b \\ &= \alpha(\nabla'_a - \nabla_a)\mu_b + (\nabla'_a - \nabla_a)\nu_b\end{aligned}$$

where, in the first step, we have used DO2, and in the second, DO4. Thus, the left side of

$$(\nabla'_a - \nabla_a)\mu_b - \gamma^m{}_{ab}\mu_m \tag{45}$$

is linear in μ_b. Therefore, there exists a tensor field $\gamma^m{}_{ab}$ such that Eqn. (45) holds for all μ_b. In particular, setting $\mu_b = \nabla_b f = \nabla'_b f$ in (45), and antisymmetrising over "a" and "b", using DO5), we have $\gamma^m{}_{[ab]}\nabla_m f = 0$ for all f. Hence, the tensor field $\gamma^m{}_{ab}$ necessarily satisfies

$$\gamma^m{}_{[ab]} = 0, \quad \text{i..e.,} \gamma^m{}_{ab} = \gamma^m{}_{(ab)} \tag{46}$$

For arbitrary ξ^b and μ_b, we have

$$0 = (\nabla'_a - \nabla_a)(\xi^b\mu_b) = \xi^b(\nabla'_a - \nabla_a)\mu_b + \mu_b(\nabla'_a - \nabla_a)\xi^b$$
$$= \xi^b\gamma^m{}_{ab}\mu_m + \mu_m(\nabla'_a - \nabla_a)\xi^m$$

where, for the first equality, we have used DO4, for the second, DO2 and DO3, and, for third (45). Since this equation holds for all μ_m, we have

$$(\nabla'_a - \nabla_a)\xi^m = -\gamma^m{}_{ab}\xi^b \tag{47}$$

for all $\$xi^b$. Finally, for a general field, e,g., $\alpha_b{}^{cd}$,

$$0 = (\nabla'_a - \nabla_a)(\alpha_b{}^{cd}\xi^b\mu_c\nu_d) = \xi^b\mu_c\nu_d(\nabla'_a - \nabla_a)\alpha_b{}^{cd}$$
$$+ \alpha_b{}^{cd}\mu_c\nu_d(\nabla'_a - \nabla_a)\xi^b + \alpha_b{}^{cd}\xi^b\nu_d(\nabla'_a - \nabla_a)\mu_c$$
$$+ \alpha_b{}^{cd}\xi^b\mu_c(\nabla'_a - \nabla_a)\nu_d$$

Using (45), (47), and the fact that ξ^b, μ_c, and ν_d are arbitrary, we obtain

$$\nabla'_a\alpha_b{}^{cd} = \nabla_a\alpha_b{}^{cd} + \gamma^m{}_{ab}\alpha_m{}^{cd} - \gamma^c{}_{am}\alpha_b{}^{md} - \gamma^d{}_{am}\alpha_b{}^{cm} \tag{48}$$

A similar formula holds, of course, for tensor fields of arbitrary rank.

We summarize:

Theorem 10. Let ∇'_a and ∇_a be derivative operators on the manifold M. Then there exists a tensor field $\gamma^m{}_{ab} = \gamma^m{}_{(ab)}$ such that Eqn. (48) holds for any tensor field, $\alpha_{b\ldots}{}^d$. Conversely, if ∇_a is any derivative operator on M, and $\gamma^m{}_{ab} = \gamma^m{}_{(ab)}$ is a tensor field on M, then ∇'_a, defined by (48), is also a derivative operator.

The first statement of Theorem 10 was proved above. The second is easy to verify: one has to check that ∇'_a, defined by (48), satisfies DO1 -DO5. [Exercise 40. Carry out this verification.] [Remark: The situation with regard to torsion should now be clear. The only place we used DO5 was to obtain (46). Thus, there is a one-to-one correspondence between operators which satisfy DO1-DO4 and pairs $(\nabla_a, \delta^m{}_{ab})$, where $\delta^m{}_{ab} = \delta^m{}_{[ab]}$, and ∇_a is a derivative operator. The dropping of DO5 merely permits one to "hide" an extra tensor field $\delta^m{}_{ab} = \delta^m{}_{[ab]}$ in his derivative operators.]

Theorem 10 is the whole story concerning uniqueness of derivative operators. A derivative operator is never unique (except when M is zero-dimensional): one can choose any old tensor field $\gamma^m{}_{ab} = \gamma^m{}_{(ab)} \neq 0$ and get another one. If one knows one derivative operator on a manifold, he knows them all. [underlineExercise 41. Write down the most general derivative operator on R^n.] The collection of all derivative operators on a manifold practically have the structure of a vector space, except that there is no natural "origin". (That is, there is no natural "zero derivative operator". If one randomly chooses some derivative operator, ∇_a, then, by Theorem 10, there is one-to-one correspondence between the collection of derivative operators on M and the vector space of tensor fields $\gamma^m{}_{ab} = \gamma^m{}_{(ab)}$.) The collection of all derivative operators on a manifold are a beautiful example of what is called an affine space.

We next consider existence of derivative operator on a manifold. The situation can be summarized by the following:

Theorem 11. A necessary and sufficient condition that a manifold M possess a derivative operator is the following: there exists a sub-collection of the collection of all the charts on M such that i) every point of M is in at least one of the U's in this sub-collection, and ii) no point of M is in an infinite number of U's in this sub-collection.

A manifold satisfying the conditions of Theorem 11 is said to be *paracompact*. [Paracompactness is actually a topological property which, in order to avoid topological questions, we have reformulated in terms of chart.] Every manifold one is ever likely to see is paracompact. In fact, it is a rather subtle business to even contract a manifold which is not paracompact. It is immediate from the definitions that the manifolds R^n and S^n are paracompact. [Exercise 42. Prove that the product of two paracompact manifolds paracompact. Exercise 43. Prove that the result of "cutting out" a closed region from a paracompact manifold is a paracompact manifold.] Although the proof of Theorem 11 is not very difficult, it is rather technical. Since neither the statement nor the proof of Theorem 11 is very important for differential geometry, we omit the proof.

Thus, every reasonable manifold has plenty of derivative operators, with any one easily obtainable from any other. A few pathological manifolds have no derivative operators.

Exercise 44. Introduce a chart. express the action of a derivative operator, ∇_a, in terms of an expression involving the components of the tensor field to which ∇_a is being applied. There will appear in your expression functions $\Gamma^i_{jk}(x)$, $i, j = 1.2, \ldots, n$. These functions are called the connection. Derive the (well-known) formula for the behavior of the connection under a coordinate transformation.

Exercise 45. Prove that every manifold has a derivative operator locally. More precisely, prove that, if p is any point of a manifold M, then there is an open subset O of M, containing p, such that the manifold $O = M - (M - O)$ possesses a derivative operator. (Thus, paracompactness is a "global property" of manifolds.)

Exercise 46. Why do not we introduce a "derivative operator" with a contravariant rather than a covariant index?

Exercise 47. Find a derivative operator on S^n.

Exercise 48. the derivative operator of Example 29 has the property that derivatives commute on an arbitrary tensor field. Verify that this property does not hold for every derivative operator on the manifold R^n.

Exercise 49 Let $\widetilde{\nabla}_a$ be an operator, defined only on covariant vector fields, which satisfies i) $\widetilde{\nabla}_a(\mu_b + \nu_b) - \widetilde{\nabla}_a\mu_b + \widetilde{\nabla}_a\nu_b$, ii) $\widetilde{\nabla}_a(\alpha\mu_b) = \alpha\widetilde{\nabla}_a\mu_b + \mu_b D(\alpha)_a$, iii) $\widetilde{\nabla}_{[a}D(\alpha)_{b]} = 0$. Prove that there exists precisely one derivative operator ∇_a which, on covariant vector fields, coincides with $\widetilde{\nabla}_a$.

11. Concomitants

Fix a derivative operator ∇_a and a vector field ξ^a. Then, from DO4, we have

$$\mathscr{L}_\xi \alpha = \xi^a \nabla_a \alpha \tag{49}$$

for every scalar field α. Now let η^a be another contravariant vector field. Then

$$\begin{aligned}\mathscr{L}_{(\mathscr{L}_\xi \eta)}\alpha &= \mathscr{L}_\xi \mathscr{L}_\eta \alpha - \mathscr{L}_\eta \mathscr{L}_\xi \alpha = \xi^b \nabla_b (\eta^a \nabla_a \alpha) - \eta^b \nabla_b (\xi^a \nabla_a \alpha) \\ &= (\xi^b \nabla_b \eta^a) \nabla_a \alpha + \xi^b \eta^a \nabla_b \nabla_a \alpha - (\eta^b \nabla_b \xi^a) \nabla_a \alpha - \eta^b \xi^a \nabla_b \nabla_a \alpha \\ &= (\xi^b \nabla_b \eta^a - \eta^b \nabla_b \xi^a) \nabla_a \alpha\end{aligned}$$

where we have used DO4 and DO5. Since α is arbitrary, we have

$$\mathscr{L}_\xi \eta^a = \xi^b \nabla_b \eta^a - \eta^b \nabla_b \xi^a \tag{50}$$

Now let μ_a be a covariant vector field. Then

$$\begin{aligned}\eta^a \mathscr{L}_\xi \mu_a &= \mathscr{L}_\xi (\mu_a \eta^a) - \mu_a \mathscr{L}_\xi \eta^a = \xi^b \nabla_b (\mu_a \eta^a) \\ &- \mu_a (\xi^b \nabla_b \eta^a - \eta^b \nabla_b \xi^a) = \eta^a (\xi^b \nabla_b \mu_a + \mu_b \nabla_a \xi^b)\end{aligned}$$

Since η^a is arbitrary, we have

$$\mathscr{L}_\xi \mu_a = \xi^b \nabla_b \mu_a + \mu_b \nabla_a \xi^b \tag{51}$$

Finally, consider an arbitrary tensor field, e.g., $\alpha^{ab}{}_{cd}$. Then

$$\begin{aligned}\mu_a \nu_b \eta^c \tau^d \mathscr{L}_\xi \alpha^{ab}{}_{cd} &= \mathscr{L}_\xi (\alpha^{ab}{}_{cd} \mu_a \nu_b \eta^c \tau^d) - \alpha^{ab}{}_{cd} \nu_b \eta^c \tau^d \mathscr{L}_\xi \mu_a \\ &- \alpha^{ab}{}_{cd} \mu_a \eta^c \tau^d \mathscr{L}_\xi \nu_b - \alpha^{ab}{}_{cd} \mu_a \nu_b \tau^d \mathscr{L}_\xi \eta^c - \alpha^{ab}{}_{cd} \mu_a \nu_b \eta^c \mathscr{L}_\xi \tau^d\end{aligned}$$

Substituting (50) and (51), and using the fact that μ_a, ν_b, $\eta^c$and τ^d are arbitrary, we have

$$\mathscr{L}_\xi \alpha^{ab}{}_{cd} = \xi^m \nabla_m \alpha^{ab}{}_{cd} - \alpha^{mb}{}_{cd} \nabla_m \xi^a - \alpha^{am}{}_{cd} \nabla_m \xi^b + \alpha^{ab}{}_{md} \nabla_c \xi^m + \alpha^{ab}{}_{cm} \nabla_d \xi^m \tag{52}$$

Thus, using an arbitrary derivative operator ∇_a, we have obtained an "explicit" expression for the Lie derivative.

It is clear from Eqn. (52) that the right side of this equation must be independent of the choice of the derivative operator ∇_a (for the left side certainly is). We verify this explicitly, using Eqn. (48), for the simpler case in which $\alpha^a{}_b$ has only two indices:

$$\begin{aligned}&\xi^m \nabla'_m \alpha^a{}_b - \alpha^m{}_b \nabla'_m \xi^a + \alpha^a{}_m \nabla'_b \xi^m = (\xi^m \nabla_m \alpha^a{}_b - \xi^m \gamma^a{}_{mn} \alpha^n{}_b \\ &+ \xi^m \gamma^n{}_{mb} \alpha^a{}_n) - (\alpha^m{}_b \nabla_m \xi^a - \alpha^m{}_b \gamma^a{}_{mn} \xi^n) + (\alpha^a{}_m \nabla_b \xi^m - \alpha^a{}_m \gamma^m{}_{bn} \xi^n) \\ &= \xi^m \nabla_m \alpha^a{}_b - \alpha^m{}_b \nabla_m \xi^a + \alpha^a{}_m \nabla_b \xi^m\end{aligned}$$

An expression involving some tensor fields and a derivative operator ∇_a, which is independent of the choice of derivative operator, is called a *concomitant*. Thus, the right side of (52) is a concomitant.

Let $\omega_{a\ldots c}$ be antisymmetric in all indices. We show that the expression

$$\nabla_{[m} \omega_{a\ldots c]} \tag{53}$$

is a concomitant. We first use (48):

$$\nabla'_m \omega_{a\ldots c} = \nabla_m \omega_{a\ldots c} + \gamma^n{}_{ma} \omega_{n\ldots c} + \cdots + \gamma^n{}_{mc} \omega_{a\ldots n}$$

Antisymmetrizing over "$ma\dots c$", and using (46), the result follows immediately. We now claim that the concomitant (53) is precisely the exterior derivative, i.e., that

$$D(\omega)_{ma\dots c} = \nabla_{[m}\omega_{a\dots c]} \tag{54}$$

For ω_a a scalar field, this is clear. For ω_a a vector field, set $\lambda = \omega_a\xi^a$. Then

$$\xi^m D(\omega)_{ma} = \frac{1}{2}\mathscr{L}_\xi\omega_a - \frac{1}{2}D(\lambda)_a = \frac{1}{2}(\xi^m\nabla_m\omega_a + \omega_m\nabla_a\xi^m)$$
$$-\frac{1}{2}\nabla_a(\xi^m\omega_m) = \frac{1}{2}\xi^m\nabla_m\omega_a - \frac{1}{2}\xi^m\nabla_a\omega_m = \xi^m\nabla_{[m}\omega_{a]}$$

Similarly for higher rank.

In Sections 7 and 9, we defined the Lie derivative and exterior derivative "algebraically". We now have expressions for these things as concomitants. We could just as well have defined the Lie derivative by (52) and the exterior derivative by (54) (after observing that the right sides of these expressions are concomitants). The various properties of the Lie and exterior derivatives are, as a rule, somewhat easier to derive from (52) and (54) than from the definitions in Sects. 7 and 9. Properties LD1, LD2, LD3, LD4, and Eqn. (29) follow from Eqn. (52) by inspection. Eqn. (31) for Lie derivatives follows from (52) after a few lines of calculation. (Exercise 50. Check this.) Properties ED1 and ED2 follow from (54) by inspection, while ELD1 and ELD2 require a few lines of calculation, and ED3 is a bit tricker (although ED3, too, will become simple after the following section). (Exercise 51. Verify ELD1 and ELD2 from (52) and (54).)

Exercise 52. Using the remarks above, together with Example 29, write down the answers to Exercises 34 and 39. (No calculations whatever are required.)

Concomitants often provide a "brute-force" way of doing things.

Example 30. We give another example of a concomitant. Let $\alpha^{a_1\dots a_p}$ and $\beta^{a_1\dots a_q}$ be symmetric in all p indices and all q indices, respectively. Set

$$[\alpha,\beta] = p\,\alpha^{m(a_1\dots a_{p-1}}\nabla_m\beta^{a_p\dots a_{p+q-1})} - q\beta^{m(a_1\dots a_{q-1}}\nabla_m\alpha^{a_q\dots a_{p+q-1})} \tag{55}$$

(Why this notation? What is (55) when $p = q = 1$?) The right side of (55) is a concomitant:

$$p\,\alpha^{ma_1\dots a_{p-1}}\nabla'_m\beta^{a_p\dots a_{p+q-1}} - q\beta^{ma_1\dots a_{q-1}}\nabla'_m\alpha^{a_q\dots a_{p+q-1}}$$
$$= p\,\alpha^{ma_1\dots a_{p-1}}(\nabla_m\beta^{a_p\dots a_{p+q-1}} - \gamma^{ap}_{mn}\beta^{n\dots a_{p+q-1}} - \dots - \gamma^{a_{p+q-1}}_{mn}\beta^{a_p\dots n})$$
$$- q\beta^{ma_1\dots a_{q-1}}(\nabla_m\alpha^{a_q\dots a_{p+q-1}} - \gamma^{a_q}_{mn}\alpha^{n\dots a_{p+q-1}} - \dots - \gamma^{a_{p+q-1}}_{mn}\alpha^{a_q\dots n})$$

Symmetrizing over the indices "$a_1\dots a_{p+q-1}$", we see that all the γ^m_{ab}−terms, cancel out. (A total of q terms arise from $\nabla_m\beta^{\cdots}$, and p terms from $\nabla_m\alpha^{\cdots}$. All these terms are equal, so they all cancel.) This concomitant has some pretty properties. In fact it satisfies (24) and (25). (Thus, we have a Lie algebra of totally symmetric contravariant tensor fields.) There is also a "Leibnitz rule". We define the only kind of "outer product" we can with these tensor:

$$\alpha\cap\beta = \alpha^{(a_1\dots a_p}\beta^{a_{p+1}\dots a_{p+q+1})} \tag{56}$$

Then

$$(\alpha,\beta\cap\gamma) = (\alpha,\beta)\cap\gamma + (\alpha,\gamma)\cap\beta \tag{57}$$

(Exercise 53. Verify that the concomitant (55) satisfies (24), (25), and (57).)

Exercise 54. Let $\alpha^{a_1\dots a_p}$ and $\beta^{a_1\dots a_q}$ be antisymmetric in all p indices and all q indices, respectively. Prove that the expression

$$p(-1)^{p+1}\alpha^{m[a_1\dots a_{p-1}}\nabla_m\beta^{a_p\dots a_{p+q-1}]} - q(-1)^{p+pq}\beta^{m[a_1\dots a_{q-1}} - \nabla_m\alpha^{a_q\dots a_{p+q-1}]} \tag{58}$$

is a concomitant.

Exercise 55. Try to make a natural-looking algebra out of the concomitant of Exercise 54, in the same way as was done in Example 30.

Exercise 56. Let $\omega_{a\ldots c}$ be a tensor field (not necessarily antisymmetric) Is $\nabla_{[m}\omega_{a\ldots c]}$ necessarily a concomitant?

Although by far the most important concomitants are the Lie derivative and the exterior derivative, the other two described above are also useful for certain applications. Although many concomitants have been written down, there is, as far as I am aware, no systematic procedure known for finding them all.

12. Curvature

The natural thing to do with derivatives is commute them. By commuting derivative operators, we are led to curvature.

Let ∇_a be a derivative operator. Then

$$\begin{aligned}&\nabla_{[a}\nabla_{b]}\left[\alpha\mu_c + \nu_c\right] = \nabla_{[a}\left[\alpha\nabla_{b]}\mu_c + (\nabla_{b]}\alpha)\mu_c + \nabla_{b]}\nu_c\right] = (\nabla_{a[}\alpha)\nabla_{b]}\mu_c\\ &+ \alpha\nabla_{[a}\nabla_{b]}\mu_c + \mu_c\nabla_{[a}\nabla_{b]}\alpha + (\nabla_{[b}\alpha)\nabla_{a]}\mu_c + \nabla_{[a}\nabla_{b]}\nu_c\\ &= \alpha\nabla_{[a}\nabla_{b]}\mu_c + \nabla_{[a}\nabla_{b]}\nu_c\end{aligned}$$

that is to say, the left side of

$$\nabla_{[a}\nabla_{b]}\mu_c = \frac{1}{2}R_{abc}{}^{d}\mu_d \tag{59}$$

is linear in μ_c. Hence, there exists a tensor field, $R_{abc}{}^{d}$, such that (59) holds for every μ_c. Furthermore,

$$\begin{aligned}&0 = \nabla_{[a}\nabla_{b]}\left[\xi^c\mu_c\right] = \nabla_{[a}\left[(\nabla_{b]}\xi^c)\mu_c + (\nabla_{b]}\mu_c)\xi^c\right] = \mu_c\nabla_{[a}\nabla_{b]}\xi^c\\ &+ (\nabla_{[b}\xi^c)\nabla_{a]}\mu_c + \xi^c\nabla_{[a}\nabla_{b]}\mu_c + (\nabla_{[a}\xi^c)\nabla_{b]}\mu^c\\ &= \mu_c\nabla_{[a}\nabla_{b]}\xi^c + \xi^c\nabla_{[a}\nabla_{b]}\mu_c\end{aligned}$$

Substituting (59) into the last expression, and using the fact that μ_c is arbitrary.

$$\nabla_{[a}\nabla_{b]}\xi^c = -\frac{1}{2}R_{abd}{}^{c}\xi^d \tag{60}$$

for every ξ^c. By the standard argument, we have, for a general tensor field, e.g., $\alpha^{cd}{}_{ps}$,

$$\nabla_{[a}\nabla_{b]}\alpha^{cd}{}_{ps} = -\frac{1}{2}R_{abm}{}^{c}\alpha^{md}{}_{ps} - \frac{1}{2}R_{abm}{}^{d}\alpha^{cm}{}_{ps} + \frac{1}{2}R_{abp}{}^{m}\alpha^{cd}{}_{ms} + \frac{1}{2}R_{abc}{}^{m}\alpha^{cd}{}_{pm} \tag{61}$$

The tensor field $R_{abc}{}^{d}$ is called the *Riemann tensor* (or curvature tensor) associated with the derivative operator ∇_a. It is immediate from the definition, e.g., (59) that

$$R_{abc}{}^{d} = R_{[ab]c}{}^{d} \tag{62}$$

There is one further algebraic condition on the Riemann tensor. Antisymmetrizing (59) over "a, b, c", we have $\nabla_{[a}\nabla_b\mu_{c]} = \frac{1}{2}R_{[abc]}{}^{d}\mu_d$ But Eqn. (54) and ED3 imply that the left side of this equation vanishes. Hence, since μ_c is arbitrary,

$$R_{[abc]}{}^{d} = 0 \tag{63}$$

Eqn. (62) and (63) are the only algebraic conditions on the Riemann tensor of a general derivative operator. There is, however, an additional differential condition. Using (59),

$$\begin{aligned}\nabla_a\nabla_{[b}\nabla_{c]}\mu_d &= \frac{1}{2}\nabla_a(R_{bcd}{}^{m}\mu_m)\\ &= \frac{1}{2}(\nabla_a R_{bcd}{}^{m})\mu_m + \frac{1}{2}R_{bcd}{}^{m}\nabla_a\mu_m\end{aligned}$$

Using (61),

$$\nabla_{[a}\nabla_{b]}\nabla_c\mu_d = \frac{1}{2}R_{abc}{}^{m}\nabla_m\mu_d + \frac{1}{2}R_{abd}{}^{m}\nabla_c\mu_m$$

Now antisymmetrize these two equations over "a, b, c". The left sides become equal; equate the right sides. Then two terms cancel, while another is eliminated by (63). Thus, we are left with $\frac{1}{2}(\nabla_{[a}R_{bc]d}{}^{m})\mu_m = 0$. Since μ_m is arbitrary,

$$\nabla_{[a}R_{bc]d}{}^{m} = 0 \tag{64}$$

Eqn. (64) is called the *Bianchi identity*.

To summarize, given any derivative operator on a manifold, there exists a tensor field field $R_{abc}{}^{d}$ such that (61) holds. This Riemann tensor satisfies the algebraic equations (62) and (63), as well as the differential equation (64).

Exercise 57. Prove that Riemann tensor associated with the derivative operator of Example 29 is zero.

Example 31. Let ∇_a and ∇'_a be derivative operators, related via Eqn. (48), and let $R_{abc}{}^{d}$ and $R'_{abc}{}^{d}$ be their respective Riemann tensors. We find an expression for $R'_{abc}{}^{d}$ in terms of $R_{abc}{}^{d}$ and γ^{m}_{ab}. For arbitrary μ_c,

$$\begin{aligned}
&\nabla'_a\nabla'_b\mu_c = \nabla'_a(\nabla_b\mu_c + \gamma^{m}{}_{bc}\mu_m) = \nabla_a(\nabla_b\mu_c + \gamma^{m}{}_{bc}\mu_m)\\
&+ \gamma^{n}{}_{ab}(\nabla_n\mu_c + \gamma^{m}{}_{nc}\mu_m) + \gamma^{n}{}_{ac}(\nabla_b\mu_n + \gamma^{m}{}_{bn}\mu_m)\\
&= \nabla_a\nabla_b\mu_c + (\nabla_a\gamma^{m}{}_{bc})\mu_m + \gamma^{m}{}_{bc}\nabla_a\mu_m\\
&+ \gamma^{n}{}_{ab}(\nabla_n\mu_c + \gamma^{m}{}_{nc}\mu_m) + \gamma^{n}{}_{ac}\nabla_b\mu_n + \gamma^{n}{}_{ac}\gamma^{m}{}_{bn}\mu_m
\end{aligned}$$

Now antisymmetrize this equation over "a" and "b". The fourth term on the right vanishes by (46), while the third and fifth terms cancel. Hence,

$$\nabla'_{[a}\nabla'_{b]}\mu_c = \nabla_{[a}\nabla_{b]}\mu_c + \mu_m\nabla_{[a}\gamma_{b]}{}^{m}{}_{c} + \mu_m\gamma^{n}{}_{c[a}\gamma^{m}{}_{b]n}$$

Now use (59) and the fact that μ_m is arbitrary:

$$R'_{abc}{}^{d} = R_{abc}{}^{d} + 2\nabla_{[a}\gamma_{b]}{}^{d}{}_{c} + 2\gamma^{n}{}_{c[a}\gamma_{b]}{}^{d}{}_{n} \tag{65}$$

This is the desired result.

Exercise 58. Verify directly from the right side of (65) that $R'_{[abc]}{}^{d} = 0$.

Exercise 59. Verify directly that right side of (65) satisfies the Bianchi identity.

Exercise 60. Derive the standard textbook formula for the components of the Riemann tensors in terms of derivatives of the connection with respect to the coordinates. (Use Exercise 57 and Example 31.)

Since derivatives should be commuted, let's commute a Lie derivative and a derivative operator.

$$\begin{aligned}
&\mathscr{L}_\xi(\nabla_a\mu_b) - \nabla_a(\mathscr{L}_\xi\mu_b) = \xi^m\nabla_m\nabla_a\mu_b + (\nabla_m\mu_b)\nabla_a\xi^m + (\nabla_a\mu_m)\nabla_b\xi^m\\
&- \nabla_a(\xi^m\nabla_m\mu_b + \mu_m\nabla_b\xi^m) = \xi^m\nabla_m\nabla_a\mu_b + (\nabla_m\mu_b)\nabla_a\xi^m + (\nabla_a\mu_m)\nabla_b\xi^m\\
&- \xi^m\nabla_a\nabla_m\mu_b - (\nabla_m\mu_b)(\nabla_a\xi^m) - \mu_m\nabla_a\nabla_b\xi^m - (\nabla_a\mu_m)(\nabla_b\xi^m)\\
&= 2\xi^m\nabla_{[m}\nabla_{a]}\mu_b - \mu_m\nabla_a\nabla_b\xi^m = (R_{nab}{}^{m}\xi^n - \nabla_a\nabla_b\xi^m)\mu_m
\end{aligned}$$

Thus, $\mathscr{L}_\xi$ and ∇_a commute, on all μ_b, if and only if

$$-\nabla_a\nabla_b\xi^m + R_{nab}{}^{m}\xi^n = 0 \tag{66}$$

in analogy with (31), we may regard the left side of (66) as the "Lie derivative of ∇_a in the xi^a-direction". (A solution ξ^a of (66) is called an affine collineation.)

Exercise 61. Check that $\mathscr{L}_\xi$ and ∇_a commute when applied to an arbitrary tensor field if and only if (66) is satisfied.

Exercise 62. Let ξ^a be an affine collineation for a derivative operator ∇_a, and let ∇'_a be another derivative operator, related to ∇_a via (48). Prove that ξ^a is an affine collineation

for ∇'_a if and only if $\mathscr{L}_\xi \gamma^m{}_{ab} = 0$. (this result further strengthens the remark above concerning the interpretation of the left side of (66).)

Exercise 63. Prove that $R_{abm}{}^m = -2R_{m[ab]}{}^m$. Using this result, check that the result of antisymmetrizing (66) over "a" and "b" is an identity. Hence, Eqn. (66) is equivalent to that equation symmetrized over "a, b". (In this form, the left side of (66) has the same structure as $\gamma^m{}_{ab}$.)

Exercise 64. Using (65), verify that there are no algebraic conditions on the Riemann tensor other than (62) and (63). That is, show that, given, given a tensor $R_{abc}{}^d$ at a point p, is satisfying (62) and (63), there exists a derivative operator whose Riemann tensor at p is $R_{abc}{}^d$.

13. Metrics

Let M be a manifold. A *metric* on M consists of a tensor field g_{ab} which is symmetric (i.e., $g_{ab} - g_{(ab)}$) and invertible (i.e., there exists a tensor field g^{ab} such that $g^{ac}g_{bc} = \delta^a{}_c$) The study of manifolds with metrics is called Riemann geometry.

We verify that the inverse, g^{ab}, of a metric g_{ab} is symmetric and unique. Symmetry follows from $g^{ac}g^{bd}g_{cd} = g^{ac}\delta^b{}_c = g^{ab}$ and the observation that, since g_{cd} is symmetric, so is the left side of this equation. To prove uniqueness, let g^{ab} and g'^{ab} be two inverses for g_{ab}. Then

$$g^{ab} = g^{ac}\delta^b{}_c = g^{ac}(g'^{bd}g_{cd}) = g'^{bd}\delta^a{}_d = g'^{ab}$$

Let M be a manifold with metricg_{ab}. It is convenient to incorporate this metric into the index notation as follows. If, e.g., $\alpha_{bs}{}^{rc}{}_{d}{}^{v}$ is any tensor field on M, we set

$$\begin{aligned}
\alpha_{bs}{}^{r}{}_{cd}{}^{v} &= g_{cm}\alpha_{bs}{}^{rm}{}_{d}{}^{v} \\
\alpha^{b}{}_{s}{}^{rc}{}_{d}{}^{v} &= g^{bm}\alpha_{ms}{}^{rc}{}_{d}{}^{v} \\
\alpha_{bs}{}^{rc}{}_{dv} &= g_{vm}\alpha_{bs}{}^{rc}{}_{d}{}^{m} \\
\alpha_{bs}{}^{rcdv} &= g^{dm}\alpha_{bs}{}^{rc}{}_{m}{}^{v}
\end{aligned}$$

This operation is called *raising and lowering of indices*. Evidently, the result of first raising an index and then lowering that index is to leave the tensor invariant. Note that the result of raising or lowering an index of a tensor field depends on the choice of metric. [If two metrics appear on a manifold, one must either suspend this convention, or select one of the metrics to be the one to be used in raising and lowering.] The use of this convention has the consequence that the metric will almost never appear explicitly.

Let g_{ab} be a metric on M and p on a point of M. The metricg_{ab} is said to have *signature* $(n^+,\ n^-)$ [n^+ and n^- are non-negative integers with $n^+ + n^- = n$, the dimension of M] at p if there are n^+ vectors $\underset{1}{\xi^a}, \ldots, \underset{n^+}{\xi^a}$ at p, and n^- vectors $\underset{1}{\eta^a}, \ldots, \underset{n^-}{\eta^a}$ at p such that

$$\begin{aligned}
&\underset{i}{\xi^a}\underset{j}{\eta_a} = 0 \quad (i = 1, \ldots, n^+;\ j = 1, \ldots, n^-) \\
&\underset{i}{\xi^a}\underset{j}{\xi_a} = 0 \quad (i, j = 1, \ldots, n^+;\ i \neq j) \\
&\underset{i}{\eta^a}\underset{j}{\eta_a} = 0 \quad (i, j = 1, \ldots, n^-;\ i \neq j) \\
&\underset{i}{\xi^a}\underset{j}{\xi_a} = +1 \quad (i = j = 1, \ldots, n^+) \\
&\underset{i}{\eta^a}\underset{j}{\eta_a} = -1 \quad (i = j = 1, \ldots, n^-)
\end{aligned}$$

Introducing a chart, the components of g_{ab} at p form an $n \times n$ matrix, which is invertible. Since every such matrix can be diagonalized, with diagonal elements all either $+1$ or -1, and since the number of $+1$'s and -1's is independent of the diagonalization, it is clear that a metric g_{ab} has a unique signature at each point p. (Of course, the ξ's and η's are not unique.) If M is n-dimensional, there are $(n + 1)$ possible signature, $(0, n)$, $(1, n - 1), \ldots, (n, 0)$. Since, in terms of a chart, the components of g_{ab} are continuous, and since the matrix of components is invertible at each point, it follows that, if g_{ab} has signature $(n^+,\ n^-)$ at p, there is an open subset O of M containing p such that g_{ab} has signature $(n^+,\ n^-)$ at each point of O. Thus, the set of points of M at which g_{ab} has any given signature is open. Since the set of points at which g_{ab} has any other signature is also open, the set of points at which g_{ab} has any given signature is also closed. A manifold M is said to be *connected* if the only subsets of M which are simultaneously open and

closed are the empty set and M itself. Thus, a metric on a connected manifold M has the same signature at every point. (Exercise 65. Verify that R^n is connected, and that S^n is connected for $n \geq 1$. Exercise 66. Prove that the product of two connected manifolds M is connected. Exercise 67. Find an example of a connected manifold M and closed subset C such that $M - C$ is not connected.)

A metric g_{ab} on M is said to be *positive-definite* if g_{ab} has signature $(n,\ 0)$, *negative-definite* if signature $(0,\ n)$, and *indefinite* if some other signature, where, in each case, these terms are used only if g_{ab} has the same signature at all points of M. The term "Riemannian geometry" is sometimes used only when g_{ab} is positive-definite, and "pseudo-Riemannian geometry" otherwise.

Exercise 68. Let the metric g_{ab} on M have signature $(n^+,\ n^-)$ (everywhere). Prove that $-g_{ab}$ is also a metric on M, and that this metric has signature $(n^-,\ n^+)$.

Exercise 69. Let g_{ab} be a positive-definite metric on M, p a point of M. Prove that, for every vector ξ^a at p, $\xi^a\xi_a \geq 0$, with equality holding if and only if $\xi^a = 0$.

Exercise 70. Let g_{ab} be an indefinite metric on M, p a point of M.

Prove that there exists a nonzero vector ξ^a at p that $\xi^a\xi_a = 0$.

Exercise 71. State and prove the triangle and Schwarz inequalities for vectors at a point of a manifold with positive-definite metric.

Example 32. Let g_{ab} and ${g'}_{ab}$ be metric on M, and suppose that, for every ξ^a, $g_{ab}\xi^a\xi^b = {g'}_{ab}\xi^a\xi^b$. We show that $g_{ab} = {g'}_{ab}$. Replacing ξ^a above by $\xi^a + \eta^a$, we have

$$g_{ab}\xi^a\xi^b + 2g_{ab}\xi^a\eta^b + g_{ab}\eta^a\eta^b = {g'}_{ab}\xi^a\xi^b + 2{g'}_{ab}\xi^a\eta^b + {g'}_{ab}\eta^a\eta^b$$

But the first term on the left equals the first term on the right, and the last term on the left equals the last term on the right. Hence, $(g_{ab} - {g'}_{ab})\xi^a\eta^b = 0$ for all ξ^a, η^b. Hence, $g_{ab} = {g'}_{ab}$.

Example 33. We construct a positive-definite metric on S^n $(n > 0)$. Let U be the region $y_{n+1} > 0$, and let the coordinates be $x^1 = y_1, \ldots, x^n = y_n$. If ξ^a is a vector field on S^n, let $\xi^1(x), \ldots, \xi^n(x)$ be its components with respect to this chart, and let g_{ab} be such that $g_{ab}\xi^a\xi^b$ becomes, in this chart, the function $[(\xi^1)^2 + \ldots + (\xi^n)^2][1 - (x^1)^2 - \ldots - (x^n)^2]^{-1}$.

Exercise 72. Prove that this defines a metric on S^n.

Exercise 73. Find a positive-definite metric on R^n, an indefinite metric on R^n.

One of the most important facts in Riemannian geometry is the following:

Theorem 12. Let M be a manifold with metric g_{ab}. Then there exists precisely one derivative operator ∇_a on M such that $\nabla_a g_{bc} = 0$.

Before proving Theorem 12, we give an argument which leads to the desired conclusion under the additional assumption that there exists at least one derivative operator, ∇'_a via (48). Then

$$\nabla'_a g_{bc} = \nabla_a g_{bc} + \gamma^m{}_{ab}g_{mc} + \gamma^m{}_{ac}g_{bm}$$

We wish to choose ∇_a, i.e., choose $\gamma^m{}_{ab}$, such that $\nabla_a g_{bc} = 0$. Thus, we have only to prove that there exists precisely one tensor field $\gamma^m{}_{ab} = \gamma^m{}_{(ab)}$ on M such that

$$\nabla'_a g_{bc} = \gamma^m{}_{ab}g_{mc} + \gamma^m{}_{ac}g_{bm} \quad (= 2\gamma_{(bc)a}) \tag{67}$$

There certainly exists such a tensor field, namely

$$\gamma^m{}_{ab} = \frac{1}{2}g^{mn}(\nabla'_a g_{bn} + \nabla'_b g_{an} - \nabla'_n g_{ab}) \tag{68}$$

(Exercise 74. Substitute (68) into (67).) To see that this solution is unique, let $\tilde{\gamma}^m{}_{ab}$ be another solution of (67), and set $\mu^m{}_{ab} = \tilde{\gamma}^m{}_{ab} - \gamma^m{}_{ab}$. Then $\mu^m{}_{ab} = \mu^m{}_{(ab)}$, and from (67), $\mu_{(ma)b} = 0$. To prove that $\mu^m{}_{ab} = 0$, we "interchange indices of μ_{mab}", first the first pair, then the second, etc.:

$$\begin{aligned}\mu_{mab} &= -\mu_{amb} = -\mu_{abm} = \mu_{bam} = \mu_{bma} \\ &= -\mu_{mba} = -\mu_{mab}\end{aligned}$$

This argument is interesting because it bring out clearly why Theorem 12 is true, namely, because $\gamma^m{}_{ab} = \gamma^m{}_{(ab)}$ and $\nabla'_a g_{bc}$ have the same index structure.

Proof of Theorem 12: For a scalar field α, set $\nabla_a \alpha = D(\alpha)_a$. For a covariant vector field, μ_b, set

$$\nabla_a \mu_b = \frac{1}{2} \mathscr{L}_\mu g_{ab} + D(\mu)_{ab}$$

where $\mu^b = g^{ab}\mu_a$. By (29) and (41), this satisfies $\nabla_a(\alpha\mu_b) = \alpha\nabla_a\mu_b + \mu_b\nabla_a\alpha$. For a contravariant vector field ξ^a, set

$$\nabla_a \xi^b = g^{bm}\nabla_a(g_{mn}\xi^n)$$

Finally, for an arbitrary tensor field, e.g., $\alpha^{cd}{}_{pq}$, define its derivative by

$$\mu_c\nu_d\xi^p\eta^q\nabla_a\alpha^{cd}{}_{pq} = \nabla_a(\alpha^{cd}{}_{pq}\mu_c\nu_d\xi^p\eta^a) - \alpha^{cd}{}_{pq}\nu_d\xi^p\eta^q\nabla_a\mu_c$$
$$- \alpha^{cd}{}_{pq}\mu_c\xi^p\eta^q\nabla_a\nu_d - \alpha^{cd}{}_{pq}\mu_c\nu_d\eta^q\nabla_a\xi^p - \alpha^{cd}{}_{pq}\mu_c\nu_d\xi^p\nabla_a\eta^q$$

(Exercise 75. Verify that this ∇_a is a derivative operator, and that $\nabla_a g_{bc} = 0$.) Uniqueness follows from the observation that, if ∇'_a is a derivative operator with $\nabla'_a g_{bc} = 0$, then, from (52) and (54), the action of ∇'_a and ∇_a coincide in each of the steps above. The ∇_a of Theorem 12 is called the derivative operator defined by g_{ab}. Whenever we have a manifold with metric, and a derivative operator ∇_a appears without a statement to the contrary, it is to be assumed that ∇_a is the derivative operator of Theorem 12. As a consequence of Theorem 12, indices of tensor fields can be raised and lowered either before or after the application of ∇_a, with the same result. That is, we have

$$\nabla_a \alpha_m{}^{cd} = g^{dp}\nabla_a \alpha_m{}^c{}_p$$

Given any tensor field, e.g., $\alpha_b{}^{cd}$, we write $\nabla_a\alpha_b{}^{cd}$ for $g_{am}\nabla_m\alpha_b{}^{cd}$. That is, raising and lowering of indices extends consistently to the index on the derivative operator.

Example 34. Let g_{ab} be a metric on M, and Ω a scalar field on M which is everywhere positive. Set $g'_{ab} = \Omega^2 g_{ab}$. (Exercise 76. Prove that g'_{ab} is a metric on M. Let ∇_a and ∇'_a be the derivative operators, respectively, with respect to g_{ab} and g'_{ab}. Then ∇_a and ∇'_a are related by (48). We find an expression for $\gamma^m{}_{ab}$. We have

$$\begin{aligned} 0 = \nabla'_a g'_{mc} &= \nabla_a g'_{bc} + \gamma^m{}_{ab} g'_{mc} + \gamma^m{}_{ac} g'_{bm} \\ &= \nabla_a(\Omega^2 g_{bc}) + \Omega^2\gamma^m{}_{ab} g_{mc} + \Omega^2\gamma^m{}_{ac} g_{bm} \\ &= 2\Omega g_{bc}\nabla_a\Omega + \Omega^2\gamma^m{}_{ab} g_{mc} + \Omega^2\gamma^m{}_{ac} g_{bm} \end{aligned}$$

Solving for $\gamma^m{}_{ab}$,

$$\gamma^m{}_{ab} = -\frac{1}{2}\Omega^{-1} g^{mn}(g_{na}\nabla_a\Omega + g_{nb}\nabla_a\Omega - g_{ab}\nabla_n\Omega \tag{69}$$

The change of metric, $g'_{ab} = \Omega^2 g_{ab}$ is called a *conformal transformation*.)

Exercise 77. Combining Examples 31 and 34, verify that, under a conformal transformation, the Riemann tensors of ∇'_a and ∇_a are related by

$$R'_{abc}{}^d = R_{abc}{}^d + \Omega^{-1} g_{c[b}\nabla_{a]}\nabla^d\Omega - \Omega^{-1}\delta^d_{\ [b}\nabla_{a]}\nabla_c\Omega + 2\Omega^{-2}(\nabla_c\Omega)\delta^d_{\ [b}\nabla_{a]}\Omega$$
$$- 2\Omega^{-2}(\nabla^d\Omega)g_{c[b}\nabla_{a]}\Omega - \Omega^{-2} g_{c[a}\delta_{b]}{}^d(\nabla^m\Omega)(\nabla_m\Omega) \quad \text{(raise, lower with gab)}$$

Exercise 78. Find an example of two distinct metrics on a manifold which determine the same derivative operator. Find metrics of different signature which define the same derivative operator.

Exercise 79. Where did we use invertibility of g_{ab} in Theorem 12? Find an example of a symmetric h_{ab} (not invertible) annihilated by two distinct derivative operators.

14. Curvature Defined by a Metric

We have seen in Sect. 13 that a metric g_{ab} on a manifold leads to a unique derivative operator ∇_a, and in Sect. 12 that a derivative operator on a manifold leads to a Riemann tensor $R_{abc}{}^{d}$. Hence, a Riemann tensor is associated with any metric. In this case, the Riemann tensor has some further properties which do not hold for Riemann tensor of an arbitrary derivative operator. Furthermore, one has, in the presence of a metric, an important tool for analyzing the Riemann tensor, namely, the ability to raise and lower indices. In this section, we consider the Riemann tensor in Riemannian geometry.

Let M be a manifold, g_{ab} a metric on M, ∇_a the corresponding derivative operator, and $R_{abc}{}^{d}$ the corresponding Riemann tensor. Then

$$0 = \nabla_{[a}\nabla_{b]}g_{cd} = \frac{1}{2}R_{abc}{}^{m}g_{md} + \frac{1}{2}R_{abd}{}^{m}g_{cm} \tag{70}$$

where the first equality follows from $\nabla_a g_{bc} = 0$, and the second from Eqn. (61). Eqn. (70) merely states that $R_{abcd} = g_{md}R_{abc}{}^{m}$ is antisymmetrical in "c, d". Combining this conclusion with (62) and (63), we have the following algebraic conditions on the Riemann tensor:

$$R_{abcd} = R_{[ab][cd]} \tag{71}$$

$$R_{[abc]d} = 0 \tag{72}$$

These, in fact, exhaust the algebraic conditions on the Riemann tensor in this case. It is convenient, however, to express them in a somewhat different form. Writing (72) explicitly, using (71), we have

$$R_{abcd} + R_{bcad} + R_{cabd} = 0$$

(Exercise 80. Why are there three terms on the left instead of six as in the last equation at the top of page 31?) Using (71) again,

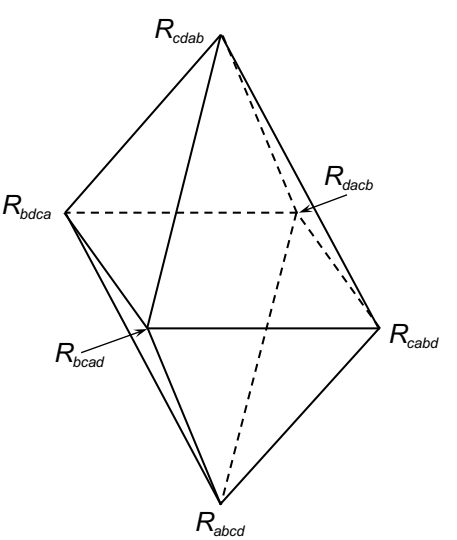
$$R_{abcd} - R_{bcda} - R_{acbd} = 0$$

Now antisymmetrize this equation over "b, c, d,". The second term on the left vanishes by (72), while the first and third terms are equal. Hence, we have

$$R_{a[bcd]} = 0 \tag{73}$$

The equation we are after is derived from (71), (72), and (73) using the figure below.

At each vertex of the octahedron, there appears a Riemann tensor with its indices in a certain order. Given any face of the octahedron, the sum of the three tensors on the vertices of that triangle vanishes, by (71), (72). For example, the sum corresponding to the rear face in the top half of the octahedron is $R_{bdca} + R_{cdab} + R_{dacb}$, which, by (71) is equal to $R_{dbac} + R_{dcba} + R_{dacb}$, which, by (73), vanishes. Now add the four expressions obtained from the four faces in the bottom half of the figure, and subtract from this the sum of the expressions obtained from the four faces in the top half of the figure. Since each



expression is zero, the result is zero. But,
when this quantity is written out, all terms
which arise from the square "equator" in the figure will cancel out. (Each such term will appear four times, twice (with plus signs) because of the sum over the top faces.) Hence, this quantity, which is zero, becomes $4R_{abcd} - 4R_{cdab}$. We conclude that

$$R_{abcd} = R_{cdab} \tag{74}$$

Eqns. (71) and (74) can now be combined into a single equation:

$$R_{abcd} = R_{[cd][ab]} \tag{75}$$

Thus, as a complete set of algebraic conditions on the Riemann tensor, we may take (72) and (75).

Example 35. Let R_{abcd} be a tensor field satisfying (75). We show that this R_{abcd} satisfies (72) if and only if it satisfies

$$R_{[abcd]} = 0 \tag{76}$$

Clearly, (72) implies (76). Next, suppose (76). Then the expression

$$R_{abcd} - R_{abdc} + R_{adbc} - R_{dabc}$$

when antisymmetrized over "a, b, c", vanishes. Using (75), this is the same as $R_{abcd} + R_{abcd} - R_{dabc} - R_{dabc}$. Hence,

$$R_{[abc]d} - R_{d[abc]} = 0$$

But, from (75), these two terms are equal. Hence, we have (72).

Example 36. We show that the complete set of algebraic conditions on the Riemann tensor can be expressed by the single condition

$$R_{abcd} = R_{[cd][ab]} + R_{[abcd]} \tag{77}$$

That is, we show that a tensor field R_{abcd} satisfies (72) and (75) if and only if it satisfies (77). Clearly, (72) and (75) imply (77). To prove the converse, antisymmetrize (77) over "a, b, c, d". All three terms become $R_{[abcd]}$, whence this quantity must vanish. Discarding this term in (77), we obtain (75). But, by Example 35, (75) and (76) imply (72).

Exercise 81. One might think he could obtain an additional algebraic condition on R_{abcd} by noting that, from Bianchi's identity,

$$\nabla_{[a}\nabla_b R_{cd]ef} = 0$$

and that, from (61), the left side can be expressed in term of the Riemann tensor with no derivatives. Check that this yields nothing new.

In Riemannian geometry, a vector field ξ^a is said to be a *Killing vector* if $\mathscr{L}_\xi g_{ab} = 0$. By (52), ξ^a is a Killing vector if and only if

$$\nabla_{(a}\xi_{b)} = 0 \tag{78}$$

The Killing vectors clearly form a vector space. By (31), they form a Lie algebra. (It is a theorem that, if M is connected, the vector space of Killing vectors on M, g_{ab} is finite -dimensional.)

Exercise 82. The remarks surrounding Eqn. (66) suggest that, if ξ^a is a Killing vector for M, g_{ab}, then ξ^a is an affine collineation for ∇_a. Prove it.

Additional curvature tensor fields can be obtained by taking contractions of the Riemann tensor. The tensor field

$$R_{ab} = R_{amb}{}^{m} \tag{79}$$

is called the *Ricci tensor* (from (75), $R_{ab} = R_{(ab)}$), and the scalar field

$$R = R_{m}{}^{m} \tag{80}$$

the *scalar curvature*. Note that the result of contracting any two indices of the Riemann tensor is either zero or $\pm R_{ab}$, and that the result of contracting all indices of the Riemann tensor is either zero or $\pm R$. Thus, the Ricci tensor and scalar curvature are the only tensor fields obtainable by contracting a single Riemann tensor. The Bianchi identity, (64), implies a differential identity on R_{ab} and R. Contracting $\nabla_{[a}R_{bc]}{}^{de}$ first over "b" and "d", and then over "c" and "e", we obtain

$$\nabla_m(R^{am} - \frac{1}{2}Rg^{am}) = 0 \tag{81}$$

One should think of the Ricci tensor and scalar curvature as "parts" of the entire Riemann tensor. This remark is made more explicit by defining the *Weyl tensor* by the equation (n = dimension $M > 2$)

$$R_{abcd} = C_{abcd} + \frac{2}{n-2}(g_{a[c}R_{d]b} - g_{b[c}R_{d]a}) - \frac{2}{(n-1)(n-2)}Rg_{a[c}g_{d]b} \tag{82}$$

Clearly, C_{abcd} has all the algebraic symmetries of the Riemann tensor. Furthermore, contracting (82), we see that all contractions of C_{abcd}, e.g., $C_{mab}{}^{m}$ vanish.

Exercise. 83 Let g_{ab} be a metric, Ω a positive scalar field, and $g'_{ab} = \Omega^2 g_{ab}$ another metric. (That is, we have a conformal transformation.) Let $C_{abc}{}^{d}$ and $C'_{abc}{}^{d}$ be the corresponding Weyl tensors. Prove that $C_{abc}{}^{d} = C'_{abc}{}^{d}$. (This fact is normally expressed by saying that the Weyl tensor is conformally invariant.)

Exercise 84. Let ξ^a be a Killing vector. Prove that $\mathscr{L}_\xi R_{abcd} = 0$. Prove, similarly, that the Lie derivative of the Ricci tensor, scalar curvature, and Weyl tensor in the $\xi-$direction are all zero.

Exercise 85. Let R_{abcd} be a tensor field satisfying (75) and (72), and set $P_{abcd} = R_{c(ab)d}$ Show that P_{abcd} satisfies

$$P_{abcd} = P_{(cd)(ab)} \quad P_{(abc)d} = 0$$

Find an expression for R_{abcd} in terms of P_{abcd}. Finally, show that, given any tensor field P_{abcd} satisfying the condition above, your expression yields an R_{abcd} satisfying (75) and (72).

15. Smooth Mappings: Action on Tensor Fields

Let M and M' be manifolds. Recall that a mapping $\psi : M \to M'$ is said to be *smooth* if, for every smooth function f' on M', the function $f = f' \cdot \psi$ on M is also smooth.

Example 37. Let $M = R^n$, $M' = R^{n'}$, and let k be a nonnegative integer less than or equal to n and less than or equal to n'. Set $\psi(x_1, \ldots, x^n) = (x^1, \ldots, x^k, 0, \ldots, 0)$. Then, if $k = 0$, all of M is mapped to the origin of M'. The smooth mapping ψ is one-to-one if and only if $k = n$, and onto if and only if $k = n'$. Finally, ψ is a diffeomorphism if and only if $k = n = n'$.

A natural question to ask is: Under what conditions does a smooth mapping from one manifold to another carry a tensor at a point, or a tensor field, form one manifold to the other? This question has a simple and complete answer, which we now obtain.

We begin with tensors at a point. Let $\psi : M \to M'$, a point of M' be smooth, let p be a point of M, and let $p' = \psi(p)$, a point of M'. Consider any contravariant vector ξ^a at the point p. Then, for any smooth function f' on M', set $\eta'(f') = \xi(f)$, where we have set $f = f' \cdot \psi$, and where $\xi(f)$ is the directional derivative as described in Sect. 3. Thus, η' associates, with any smooth function f' on M', a real number, $\eta'(f')$. We verify that this $\eta'(f')$ satisfies DD1–DD3 of Sect. 3. Properties DD1 and DD3 are obvious. To check DD2, note that, if $f' = g'h'$, then $f = gh$. Furthermore, $f'(p') = f(p)$, $g'(p') = g(p)$, and $h'(p') = h(p)$. The verification of DD2 is now easy: $\eta'(g'h') = \xi(gh) = g(p)\xi(h) + h(p)\xi(g) = g'(p')\eta'(h') + h'(p')\eta(g')$. Thus, a contravariant vector ξ at the point p of M defines a contravariant vector η' at the point p' of M'. We write $\eta'^a = \vec{\psi}_p \xi^a$. Thus, $\vec{\psi}_p$ is a (clearly) linear mapping from the vector space of contravariant vectors at p of M to the vector space of contravariant vectors at p' of M'.

Next, let μ'_a be a covariant vector in M' at p'. Then, if ξ^a is any contravariant vector in M at p, consider the number $\mu'_a(\vec{\psi}_p \xi^a)$. (This is well-defined, for $\vec{\xi}_p \xi^a$ is a contravariant vector in M' at p', while μ'_a is a contravariant vector in M' at p'.) The number $\mu'_a(\vec{\psi}_p \xi^a)$ is clearly linear in ξ^a. Hence, we have just described a linear mapping from contravariant vectors in M at p to real numbers. In other words, we have just described a covariant vector in M at p. We write this covariant vector as $\underset{\leftarrow}{\psi}_p \mu'_a$, so $\underset{\leftarrow}{\psi}_p \mu'_a$, is defined by the equation $\xi^a(\underset{\leftarrow}{\psi}_p \mu'_a) = \mu'_a(\vec{\psi}_p \xi^a)$ for all ξ^a in M at p. Thus, $\underset{\leftarrow}{\psi}_p$ is a (clearly) linear mapping from the vector space of covariant vectors in M' at p' to the vector space of covariant vectors in M at p.

Finally, we extend to tensors – rather than just vectors – at a point. Let $\alpha^{a\ldots c}$ be a contravariant tensor (= tensor all of whose indices are contravariant) in M at p. Then, for any vectors $\mu'_a, \ldots, \nu'_c$ in M' at p', the right side of

$$(\vec{\psi}_p \alpha^{a\ldots c})\mu'_a \ldots \nu'_c = \alpha^{a\ldots c}(\underset{\leftarrow}{\psi}_p \mu'_a) \ldots (\underset{\leftarrow}{\psi}_p \nu'_c) \tag{83}$$

is multilinear in $\mu'_a, \ldots, \nu'_c$. Hence, Eqn. (83) defines a contravariant tensor, $\vec{\psi}_p \alpha^{a\ldots c}$, at the point p' of M'. Similarly, if $\beta'_{a\ldots c}$ is a covariant tensor in M' at p', and $\xi^a, \ldots, \eta^c$ are vectors in M at p, then the right side of

$$(\underset{\leftarrow}{\psi}_p \beta'_{a\ldots c})\xi^a \ldots \eta^c = \beta'_{a\ldots c}(\vec{\psi}_p \xi^a) \ldots (\vec{\psi}_p \eta^c) \tag{84}$$

is multilinear in $\xi^a, \ldots, \eta^c$, defining a covariant tensor, $\underset{\leftarrow}{\psi}_p \beta'_{a\ldots c}$, in M at p. Clearly, the mappings $\vec{\psi}_p$ and $\underset{\leftarrow}{\psi}_p$, as extended above to tensor of arbitrary rank, are linear.

We now have a mapping $\vec{\psi}_p$ from contravariant tensors in M at p to contravariant tensors in M' at p', and a mapping $\underset{\leftarrow}{\psi}_p$ from covariant tensors in M' at p' to covariant tensors in M at p. How do these mappings interact with the tensor operations? We first consider outer product. Let ${\alpha'}_{a\ldots c}$ and ${\beta'}_{b\ldots d}$ be at p', and $\xi^a, \ldots, \tau^c, \eta^b, \ldots, \lambda^d$ be at p. Then,

$$\begin{aligned}
&\left[\underset{\leftarrow}{\psi}_p\,({\alpha'}_{a\ldots c}{\beta'}_{b\ldots d})\right]\xi^a\cdots\lambda^d = {\alpha'}_{a\ldots c}{\beta'}_{b\ldots d}(\vec{\psi}_p\xi^a)\ldots(\vec{\psi}_p\lambda^d)\\
&= \left[{\alpha'}_{a\ldots c}(\vec{\psi}_p\xi^a)\ldots(\vec{\psi}_p\tau^c)\right]\left[{\beta'}_{b\ldots d}(\vec{\psi}_p\eta^b)\ldots(\vec{\psi}_p\lambda^d)\right]\\
&= (\underset{\leftarrow}{\psi}_p\,{\alpha'}_{a\ldots c})\xi^a\ldots\tau^c\,(\underset{\leftarrow}{\psi}_p\,{\beta'}_{b\ldots d})\eta^b\ldots\lambda^d
\end{aligned}$$

Evidently, $\underset{\leftarrow}{\psi}_p\,({\alpha'}_{a\ldots c}{\beta'}_{b\ldots d}) = \underset{\leftarrow}{\psi}_p\,({\alpha'}_{a\ldots c})\,\underset{\leftarrow}{\psi}_p\,({\beta'}_{b\ldots d})$. Similarly for $\vec{\psi}_p$.

The second operation to be considered is contraction. Let $\alpha^{a\ldots bc\ldots d}$ be at p, and ${\mu'}_a\ldots,\kappa'_b,{\nu'}_c,\ldots,{\rho'}_d$ be at p'. Then

$$\begin{aligned}
&(\vec{\psi}_p\alpha^{a\ldots d}){\mu'}_a\ldots\kappa'_b{\nu'}_c\ldots{\rho'}_d = \alpha^{a\ldots d}(\underset{\leftarrow}{\psi}_p\,{\mu'}_a)\ldots(\underset{\leftarrow}{\psi}_p\,\kappa'_b)(\underset{\leftarrow}{\psi}_p\,{\nu'}_c)\ldots(\underset{\leftarrow}{\psi}_p\,{\rho'}_d)\\
&= \vec{\psi}_p\left[\alpha^{a\ldots d}(\underset{\leftarrow}{\psi}_p\,{\mu'}_a)\ldots(\underset{\leftarrow}{\psi}_p\,\kappa'_b)\right]{\nu'}_c\ldots{\rho'}_d
\end{aligned}$$

where the first equally is just (83), and in the second equality, we are again using (83), but taking ${\mu'}_a,\ldots,\kappa'_b$ as fixed. Since ${\nu'}_c,\ldots,{\rho'}_d$ are arbitrary, we have

$$(\vec{\psi}_p\alpha^{a\ldots d}){\mu'}_a\ldots\kappa'_b = \vec{\psi}_p\left[\alpha^{a\ldots d}\underset{\leftarrow}{\psi}_p\,{\mu'}_a)\ldots(\underset{\leftarrow}{\psi}_p\,\kappa'_b)\right]$$

Finally, noting that every covariant tensor at p' can be written as a sum outer products of vectors, we have

$$(\vec{\psi}_p\alpha^{a\ldots bc\ldots d}){\beta'}_{a\ldots b} = \vec{\psi}_p\left[\alpha^{a\ldots bc\ldots d}\underset{\leftarrow}{\psi}_p\,{\beta'}_{a\ldots b}\right] \tag{85}$$

Similarly with contravariant and covariant, p and p', etc. reversed.

We summarize:

Theorem 13. Let $\psi : M \to M'$ be a smooth mapping of manifolds. Let p be a point of M, and let $p' = \psi(p)$, a point of M'. Then there are natural, rank-preserving mappings $\vec{\psi}_p$ from contravariant tensors in M at p to contravariant tensors in M' at p', and $\underset{\leftarrow}{\psi}_p$ from covariant tensor in M' at p' to covariant tensors in M at p, such that:

1. $\vec{\psi}_p(\alpha^{a\ldots c} + c\beta^{a\ldots c}) = \vec{\psi}_p\alpha^{a\ldots c} + c\vec{\psi}_p\beta^{a\ldots c}$
   $\underset{\leftarrow}{\psi}_p\,({\gamma'}_{a\ldots c} + c{\mu'}_{a\ldots c}) = \underset{\leftarrow}{\psi}_p\,{\gamma'}_{a\ldots c} + c\underset{\leftarrow}{\psi}_p\,({\mu'}_{a\ldots c}$

where c is a constant.

2. $\vec{\psi}_p(\alpha^{a\ldots c}\beta^{b\ldots d}) = (\vec{\psi}_p(\alpha^{a\ldots c})(\vec{\psi}_p(\beta^{b\ldots s})$
   $\underset{\leftarrow}{\psi}_p\,({\mu'}_{a\ldots c}\,{\nu'}_{b\ldots d}) = \underset{\leftarrow}{\psi}_p\,({\mu'}_{a\ldots c})(\underset{\leftarrow}{\psi}_p\,({\nu'}_{b\ldots d}$

3. $(\vec{\psi}_p\alpha^{a\ldots bc\ldots d}){\mu'}_{a\ldots b} = \vec{\psi}_p\left[\alpha^{a\ldots bc\ldots d}\,\underset{\leftarrow}{\psi}_p\;{\mu'}_{a\ldots b}\right]$
   $(\underset{\leftarrow}{\psi}_p\;({\mu'}_{a\ldots bc\ldots d})\beta^{a\ldots b} = \underset{\leftarrow}{\psi}_p\left[{\nu'}_{a\ldots bc\ldots d}\vec{\psi}_p\beta^{a\ldots b}\right]$

Theorem 13 is essentially the whole story regarding tensors at a point. There are no other "natural" mappings available, i.e., no reasonable way in general to carry a covariant

tensor in M at p to a covariant tensor in M' at p'. However, when we pass to tensor fields, some more things are true.

We begin with contravariant fields on M. Let $\xi^{a\ldots c}$ be such a tensor field. Then, for each point p of M, $\vec{\psi}_p\ \xi^{a\ldots c}(p)$ is a tensor at the point $p' = \psi(p)$ of M! Does this define a tensor field on M'? The answer is no. For one thing, there may be distinct points p and q of M such that $\psi(p) = \psi(q) = p'$ (i.e., ψ may not be one-to-one; Example 37) and such that $\vec{\psi}_p\xi^{a\ldots c}(p)$ and $\vec{\psi}_q\xi^{a\ldots c}(q)$ (both tensors in M' at p') are not equal. Which tensor at p' should we choose to get a tensor field on M'? Furthermore, there may be points of M' which are not the image under of any point of M (i.e., ψ may not be onto; Example 37). A contravariant tensor field on M defines no tensor whatever at such points of M'. Clearly, this is no good.

We try it the other way. Let $\mu'_{a\ldots c}$ be a tensor field on M'. Then, for each point p of M $\underset{\leftarrow}{\psi}_p\ (\mu'_{a\ldots c}(p')$ is a covariant tensor at p. Since this is true for each point p of M, we have defined a covariant tensor field on M. Thus, with each covariant tensor field $\mu'_{a\ldots c}$ on M' we associate a covariant tensor field, $\underset{\leftarrow}{\psi}\ \mu'_{a\ldots c}$ on M. (We have not yet proven that $\underset{\leftarrow}{\psi}\ \mu'_{a\ldots c}$ is smooth if $\mu'_{a\ldots c}$ is.) Evidently, the linearity and outer-product behavior from Theorem 13 extend from tensors at a point to tensor fields.

The mapping $\underset{\leftarrow}{\psi}$ is not very interesting unless it takes smooth tensor fields to smooth tensor fields. We now show that it does. First, scalar fields (covariant tensor fields of rank zero). If f' is a smooth scalar field on M', then $\underset{\leftarrow}{\psi} f' = f = f' \cdot \psi$ is also smooth, by the definition of a smooth mapping. We next show that

$$D(\underset{\leftarrow}{\psi} f')_a = \underset{\leftarrow}{\psi}[D(f')_a] \tag{86}$$

Let ξ^a be a vector in M at p. Then

$$\begin{aligned}\xi^a D(\underset{\leftarrow}{\psi} f')_a\Big|_p &= \xi^a D(f)_a\Big|_p = \xi(f) = [\vec{\psi}_p\xi](f') \\ &= (\vec{\psi}_p\xi^a)D(f')_a\Big|_p = \xi^a \underset{\leftarrow}{\psi}\ [D(f')_a]\Big|_p\end{aligned}$$

where the second equality is the definition of the gradient, the third is the definition of $\vec{\psi}_p$, the fourth is the definition of the gradient, and the fifth is the definition of $\underset{\leftarrow}{\psi}$. That is to say, Eqn. (86), evaluated at any point p of M and contracted with any contravariant vector in M at p, holds. Hence, Eqn. (86) holds. Now let μ'_a be a covariant vector field in M'. Then, for each point p' of M', there exists an open subset O' of M', containing p', and smooth functions $f', g' \ldots, h', k'$ on M' such that, in O',

$$\mu'_a = f'\ D(g')_a + \ldots + h'\ D(k')_a \tag{87}$$

(Proof: Introduce a chart, use components, and repeat the discussion of page 20.) Hence, from (86)

$$\underset{\leftarrow}{\psi}\ \mu'_a = (\underset{\leftarrow}{\psi}\ f')\ D\ (\underset{\leftarrow}{\psi}\ g')_a + \ldots + (\underset{\leftarrow}{\psi}\ h')D\ (\underset{\leftarrow}{\psi} k')_a \tag{88}$$

Since the right side of (88) is smooth, we conclude that $\underset{\leftarrow}{\psi}\ \mu'_a$ is smooth in $\vec{\psi}^{-1}[O']$. But every point of M' is contained in such an O', and so $\underset{\leftarrow}{\psi}\ \mu'_a$ is smooth on M. Finally, let $\alpha'_{a\ldots c}$ be a tensor field on M' of arbitrary rank. Then every point of M' is contained in an open set in which

$$\alpha'_{a\ldots c} = \mu'_a \cdots k'_c + \cdots + \nu'_a \cdots \rho'_c \tag{89}$$

Then, in ψ^{-1} of this open set,

$$\underset{\leftarrow}{\psi}\ \alpha'_{a\ldots c} = (\underset{\leftarrow}{\psi}\ \mu'_a) \cdots (\underset{\leftarrow}{\psi}\ \kappa'_c) + \cdots + (\underset{\leftarrow}{\psi}\ \nu'_a) \cdots (\underset{\leftarrow}{\psi}\ \rho'_c)$$

Since the right side is smooth, so is the left, Since every point of M' is contained in such an open set, $\underset{\leftarrow}{\psi}\ \alpha'_{a\ldots c}$ is a smooth tensor field on M.

Thus, $\underset{\leftarrow}{\psi}$ is a rank-preserving mapping from smooth covariant tensor fields on M' to smooth covariant tensor fields on M. Finally, we drove that "$\underset{\leftarrow}{\psi}$ commutes with exterior differentiation:"

$$\underset{\leftarrow}{\psi}\ D(\omega')_{ma\ldots c} = D(\underset{\leftarrow}{\psi}\ \omega')_{ma\ldots c} \tag{90}$$

For ω' a 0−form, (90) is just (86). For ω' a 1−form, we take the exterior derivative of (87),

$$D(\omega')_{ma} = D(f')_{[m}D(g')_{a]} + \cdots + D(h')_{[m}D(k')_{a]}$$

and apply $\underset{\leftarrow}{\psi}$,

$$\begin{aligned}\underset{\leftarrow}{\psi}\ D(\omega')_{ma} &= D(\underset{\leftarrow}{\psi}\ f')_{[m}\ D(\underset{\leftarrow}{\psi}\ g')_{a]} + \cdots + D(\underset{\leftarrow}{\psi}\ h')_{[m}\ D(\underset{\leftarrow}{\psi}\ k')_{a]}\\ &= D(\underset{\leftarrow}{\psi}\ f'\ D(\underset{\leftarrow}{\psi}\ g') + \cdots + \underset{\leftarrow}{\psi}\ (h')\ D(\underset{\leftarrow}{\psi}\ k'))_{ma}\\ &= D(\underset{\leftarrow}{\psi}\ \omega')_{ma}\end{aligned}$$

For $\omega_{a\ldots c}$ a p−form, we take the exterior derivative of (89),

$$\begin{aligned}&D(\alpha')_{ma\ldots c} = D(\mu')_{ma}\cdots\kappa'_{\ c} + \cdots + \mu'_{\ a}\cdots D(\kappa')_{mc}\\ &+\cdots + D(\nu')_{ma}\cdots\rho'_{\ c} + \cdots + \nu'_{\ a}\cdots D(\rho')_{mc}\end{aligned}$$

and apply $\underset{\leftarrow}{\psi}$,

$$\begin{aligned}&\underset{\leftarrow}{\psi}\ D(\alpha')_{ma\ldots c} = D(\underset{\leftarrow}{\psi}\ \mu')_{ma}\cdots\underset{\leftarrow}{\psi}\ \kappa'_{\ c} + \cdots + \underset{\leftarrow}{\psi}\ \mu'_{\ a}\ldots D(\underset{\leftarrow}{\psi}\ \kappa')_{mc}\\ &+\cdots + D(\underset{\leftarrow}{\psi}\ \nu')_{ma}\cdots\underset{\leftarrow}{\psi}\ \rho'_{\ c} + \cdots + \underset{\leftarrow}{\psi}\ \nu'_{\ a}\cdots D(\underset{\leftarrow}{\psi}\ \rho')_{mc}\\ &= D(\underset{\leftarrow}{\psi}\ \alpha')_{ma\cdots c}\end{aligned}$$

To summarize,

Theorem 14. The mapping $\underset{\leftarrow}{\psi}_p$ of Theorem 13 extends to a mapping, $\underset{\leftarrow}{\psi}$ from (smooth) covariant tensor fields on M' to (smooth) covariant tensor fields on M, with $\underset{\leftarrow}{\psi}$ satisfying the following:

1. $\underset{\leftarrow}{\psi}(\gamma'_{\ a\cdots c} + \mu'_{\ a\cdots c}) = \underset{\leftarrow}{\psi}(\gamma'_{\ a\cdots c}) + \underset{\leftarrow}{\psi}(\mu'_{\ a\cdots c})$.
2. $\underset{\leftarrow}{\psi}(mu'_{\ a\cdots c}\nu'_{\ b\cdots d}) = \underset{\leftarrow}{\psi}(\mu'_{\ a\cdots c})\underset{\leftarrow}{\psi}(\nu'_{\ b\cdots d})$.
3. For$\omega_{a\cdots}$ a p−form, $\underset{\leftarrow}{\psi}\ D(\omega)_{ma\cdots} = D(\underset{\leftarrow}{\psi}\ \omega)_{ma\cdots c}$

We summarize what exists in the following table:

	Tensor at a point			Tensor fields		
$\psi : M \to M'$	contra-variant	co-variant	mixed	contra-variant	co-variant	mixed
From M to M'	$\vec{\psi}_p$	—	—	—	—	—
From M' to M	—	$\underset{\leftarrow}{\psi}_p$	—	—	$\underset{\leftarrow}{\psi}$	—

Finally, suppose that $\psi : M \to M'$ is a diffeomorphism, so $\psi^{-1} : M' \to M$ also exists and is smooth. Let $\alpha'^{a\cdots c}_{\ \ \ \ b\cdots d}$ be a tensor field on M'. Then, for any vector fields $\mu_a \cdots \nu_c$

on M, the right side of

$$\beta^{a\cdots c}{}_{b\cdots d}\,\mu_a\cdots\nu_c = \underset{\leftarrow}{\psi}\left[\alpha'^{a\cdots c}{}_{b\cdots d}\,(\underset{\leftarrow}{\psi}^{-1}\mu_a)\cdots(\underset{\leftarrow}{\psi}^{-1}\nu_c)\right]$$

is linear in $\mu_a\cdots\nu_c$. Hence, we define a tensor field, $\beta^{a\cdots c}{}_{b\cdots d}$ on M. Clearly, a diffeomorphism between two manifolds defines a one-to-one correspondence between the tensor fields on one manifold and those on the other, where this correspondence preserves everything (rank, symmetry, number of contravariant and covariant indices, commutes with addition, outer product, index substitution, and contraction, commutes with Lie and exterior differentiation, etc.). This is to be expected. A diffeomorphism between manifolds means they are "identical as manifolds", while tensor fields exist on things with a manifold structure.

Exercise 86. Determine $\vec{\psi}_p$ and $\underset{\leftarrow}{\psi}_p$ for the mappings of Example 37.

Exercise 87. Why don't we write $\underset{\leftarrow}{\psi}_{p'}$ instead of $\underset{\leftarrow}{\psi}_p$?

Exercise 88. Why don't we ask what happens to Lie derivatives under $\underset{\leftarrow}{\psi}$?

Exercise 89. Find an example of a smooth onto mapping ψ such that $\vec{\psi}_p$ is an isomorphism (of vector spaces) for every p, but such that ψ is not a diffeomorphism.

Exercise 90. Let $\psi : M \to M'$ be smooth, so $\varphi\cdot\psi : M' \to M''$ be smooth, so $\varphi\cdot\psi : M \to M''$ is smooth. Prove that $\underset{\leftarrow}{|\psi|}\cdot\underset{\leftarrow}{|\varphi|} = \underset{\leftarrow}{(\varphi\cdot\psi)}$. Let p be a point of M, $p' = \psi(p)$, and $p'' = \varphi(p')$. Prove that $\underset{\leftarrow}{(\varphi\cdot\psi)}_p = \underset{\leftarrow}{\psi}_p\cdot\underset{\leftarrow}{\varphi}_{p'}$ and $\vec{(\varphi\cdot\psi)}_p = \vec{\phi}_{p'}\cdot\vec{\psi}_p$.

Exercise 90a. Find an example in which g'_{ab} is a metric on M', but $\underset{\leftarrow}{\psi}\, g'_{ab}$ is not a metric on M.

Exercise 91. Let $\psi : M \to M'$ be smooth. Show that there is no natural way, in general, to take a derivative operator from M to M', or to take a derivative from M' to M.

Exercise 92. Let $\psi : M \to M'$ be smooth. Let ξ^a be a vector field on M such that, for every point p of M, $\vec{\psi}_p\,\xi^a(p) = 0$. Prove that, for every $\alpha'_{a\cdots c}$ on M', $\mathscr{Z}_\xi\,(\underset{\leftarrow}{\psi}\,\alpha'_{a\cdots c}) = 0$

Exercise 93. Prove that $\vec{\psi}_p$ (resp. $\underset{\leftarrow}{\psi}_p$) is onto if and only if $\underset{\leftarrow}{\psi}_p$ (resp. $\vec{\psi}_p$) is one-to-one.

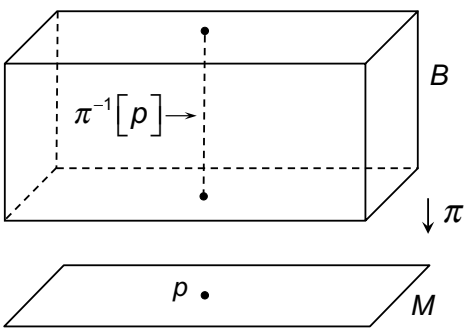
Exercise 94. Let g_{ab} be a metric on M, and g'_{ab} a metric on M'. A diffeomorphism $\psi : M \to M'$ is said to be an *isometry* if $\underset{\leftarrow}{\psi}\, g'_{ab} = g_{ab}$. Prove that, if ψ is an isometry, then $\underset{\leftarrow}{\psi}\, R'_{abcd} = R_{abcd}$.

Exercise 95. Express the action of $\vec{\psi}_p$ in terms of components with respect to charts. (The mapping $\vec{\psi}_p$ is often called the derivative of ψ at p.)

16. Bundles

We have seen several situations in which, at each point of a manifold, there sits a "space". For example, at each point p of a manifold, we have the space of contravariant vectors at p. There exists a class of mathematical objects, called fibre bundles, in which structure of this general type is isolated. A certain class of fibre bundles, the smooth vector bundles, are rather simpler to define, and, furthermore, suffice for essentially all applications in differential geometry. In this section, we define a smooth vector bundle, and derive a few of its elementary properties.

A smooth vector bundle consists, firstly, of a manifold B (called the *bundle space*), a manifold M (called the *base space*), and a smooth mapping $\pi : B \to M$ (called the *projection*). For each point p of M, the subset $\pi^{-1}[p]$ (i.e, the set of all points P of B such that $\pi(P) = p$) of B is called the *fibre* over p. It is convenient to represent this situation by a figure with the bundle space drawn over the base space, and the projection mapping taking each point of B "vertically downward" to the point of M directly below it:



In terms of the figure, the fibre over p is the vertical line in B directly above p. We next require that each fibre have the structure of a $k-$dimensional vector space, where the non-negative integer k is fixed (i.e., does not vary from one fibre to another). In more detail, this means that linear combinations (with real coefficients) of elements of a single fibre are defined, that these combinations are again elements of that fibre, and that all the axioms for a $k-$dimensional vector space are satisfied by this operation on this one fibre. (We cannot "add" elements of different fibres.)

This is not quite the end of the definition. We still wish to ensure that "locally, the bundle space B is a product of the base space and a $k-$dimensional vector space." This idea is formulated as follows. Let U be a "sufficiently small" open subset of M. Then $\pi^{-1}[U]$ (i.e., the subset of B consisting of all points of B which are mapped into U by π) should have "exactly the same structure" as $U \times V^k$, where V^k is a $k-$dimensional vector space (and, therefore, a $k-$dimensional manifold, so $U \times V^k$ is a product of manifolds). That is, we require that there exists a diffeomorphism $\psi : U \times V^k \to \pi^{-1}[U]$ which is "structure preserving". What is there to be preserved? For each point p of U, denote by $\{p\}$ the subset of $U \times V^k$ consisting of all pairs (p, v), for all elements v of V^k. We require, firstly, that ψ map each subset of $U \times V^k$ of the form $\{p\}$ onto the fibre over p. In other words, ψ should be "fibre preserving". That is, we require that the point $\psi(p, v)$ of B lie in the fibre over p. That is, we require that $\pi \cdot \psi(p, v) = p$. There is one more piece of structure to be preserved by ψ. Each subset $\{p\}$ of $U \times V^k$ is, of course, a $k-$dimensional vector space. But each fibre in B is also a $k-$dimensional vector space. Since ψ takes $\{p\}$ to the fibre over p, we might as well require that that ψ also preserve the vector-space structures. Explicitly, we require that, for v and w in V^k and c a number, $\psi(p, v + cw) = \psi(p, v) + c\,\psi(p, w)$. (Note that the right side is well-defined, for $\psi(p, v)$ and $\psi(p, w)$ lie in the same fibre, namely, the fibre over p.) Thus, the "local product" condition is that, for every point q of M, there exists an open subset U of M and a diffeomorphism $\psi : U \times V^k \to \pi^{-1}[U]$ which is fibre-preserving and vector-space structure preserving.

We summarize this discussion with the definition. A (*smooth*) *vector bundle* consists of a manifold B, a manifold M, a smooth mapping $\pi : B \to M$, a non-negative integer

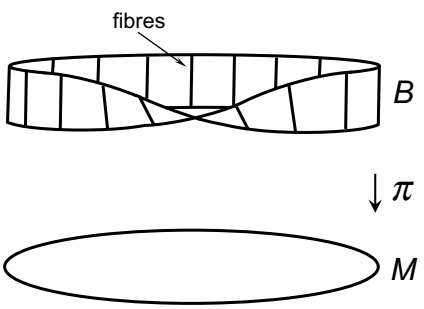
k, and a $k-$dimensional vector space structure on $\pi^{-1}[p]$ for each point p of M such that: for each point q of M there is an open subset U of M, containing q, and a diffeomorphism $\psi : U \times V^k \to \pi^{-1}[U]$ satisfying i) $\pi \cdot \psi(p, v) = p$, and ii) $\psi(p, v + cw) = \psi(p, v) + c\psi(p, w)$.

Example 38. Let M be a manifold, and V^k a $k-$dimensional vector space. Set $B = M \times V^k$, and let $\pi : B \to M$ be defined by $\pi(p, v) = p$. Then the fibre over a point p of M is the collection of all points (p, v) of B for all v in V^k. Clearly, these fibres have the structure of a $k-$dimensional vector space. It is also clear that this is a fibre bundle. (Not only is B a "local product", but also a "global product".)

Example 39. Let $M = R^1$, so a point of M is a real number x. Let $B = R^2$, so a point of B is a pair, (y, z), of real numbers. Let π be defined by $\pi(y, z) = y^3$, a point of M. Then the fibre over the point x of M is the subset of B consisting of pairs $(x^{1/3}, z)$, for all z. These fibres have an obvious one-dimensional vector space structure, i.e., $(x^{1/3}, z) + c(x^{1/3}, w) = (x^1/3, z + cw)$. However, this is not a vector bundle, for B is not a local product near the point $x = 0$ of M. (Exercise 96. Verify that the above structure does not define a vector bundle.)

We now obtain a few properties of vector bundles. Note, firstly, that the projection π must be onto M, for, if a point q of M were not in the image of B under π, then $\pi^{-1}[q]$ would be the empty set. But the empty set cannot have the structure of a $k-$dimensional vector space. Next, note, that, if the base space M is $n-$dimensional, and if the fibres are $k-$dimensional vector spaces, then the bundle space B must have dimension $(n + k)$. This is immediate from the local product property and the fact that, for products of manifolds, the dimensions add. (That is, $U \times V^k$ has dimension $(n + k)$.) The next property is slightly more subtle. Let P be a point of B, and let $\pi(P) = p$, a point of M. Then $\vec{\pi}_P$ is a linear mapping from the $(n + k)-$dimensional vector space of contravariant vectors in M at p. We want to describe this mapping $\vec{\pi}_P$ in more detail. Let U and $\psi : U \times V^k \to \pi - 1[U]$ be as in the local product part of the definition, and let v be the element of V^k such that $\psi(p, v) = P$. Let $\alpha : U \to U \times V^k$ be defined by $\alpha(q) = (q, v)$, and $\beta : V^k \to U \times V^k$ defined by $\beta(w) = (p, w)$. Then $\psi \cdot \beta : V^k \to B$ maps V^k to the fiber over p. Hence, $\pi \cdot \psi \cdot \beta : V^k \to M$ maps all of V^k to the single point of p of M. Hence, any contravariant vector in the manifold V^k at the point v is mapped, by $(\pi \cdot \vec{\psi} \cdot \beta)_v$ to the zero contravariant vector in M at p. On the other hand, $\pi \cdot \psi \cdot \alpha : U \to U$ is just the identity mapping. Hence, $(\pi \cdot \vec{\psi} \cdot \alpha)_p$ maps a contravariant vector in U at p to that same contravariant vector in U at p. Thus, the $(n + k)-$dimensional vector space of contravariant vectors in B at P has a $k-$dimensional subspace consisting of vectors which get mapped to zero, and a $n-$dimensional subspace consisting of vectors which (except for the zero one) do not get mapped to zero. That is, $\vec{\pi}_P$, a linear mapping from a $(n + k)-$dimensional space to an $n-$dimensional space, is onto, and annihilates a certain $k-$dimensional subspace of the vector space of contravariant vectors in B at P. Since vector bundles are "locally products", things like $\vec{\pi}_P$, which are local, have exactly the same properties that they have for products.

Example 40. If, in Example 38, we let $M = S^1$ (the circle), and $k = 1$ (so V^k is a one-dimensional vector space, i.e., a line), then the bundle space B is, a manifold, the cylinder, $S^1 \times R^1$. But there is another bundle with base space S^1, and with $k = 1$, which is not a simple product. Let B consist of pairs, (θ, x), of real numbers, where the pair (θ, x) and $(\theta + n\,2\pi, \pm x)$ are identified, with the plus sign if n is an even integer, minus sign if an odd integer. Let M consist of numbers θ, with θ and $\theta + n\,2\pi$ (n an integer) identified. Thus, $M = S^1$ (setting $\tan\theta = y_2/y_1$ to obtain the usual representation of S^1). Set $\pi(\theta, x) = \theta$. The fibre over a point θ of M consists of all pairs, (θ, x), with vector space



structure $(\theta, x) + c(\theta, y) = (\theta, x + cy)$. This is a vector bundle. Of course, the bundle space B is the Mobious strip.

Why bother with these vector bundles and "local product structures" instead of just considering products as in Example 38? Vector bundles have two nice properties that products do not. Firstly, Example 40 shows that vector bundles can have interesting global structure not possible for a product. Most vector bundles in practice are not products. Secondly, and more important, vector bundles have certain of the structure of a product "washed out". In the product, $M \times V^k$, one can say that two points, (p, y) and (q, w), lie "at the same places on their respective fibres" if $v = w$ (elements of V^k). In other words, there are natural isomorphisms between all the fibres. This is structure which is not present in a general vector bundle. In general, one cannot compare points on different fibres. This is what is wanted for applications.

Exercise 97. Define an isomorphism of vector bundles.

Exercise 98. Prove that B is connected if and only if M is.

Exercise 99. Prove that every vector bundle with base space $M = R^1$ is a product bundle.

Exercise 100. Prove that S^2 is not the bundle space for any vector bundle.

Exercise 101. Consider two vector bundles with the same base space M, and with fibres of dimension k and k'. By taking the direct sum (of vector spaces) within each fibre, obtain a third vector bundle with base space M, and fibres of dimension $(k + k')$.

17. The Tensor Bundles

With any manifold M there is associated a collection of vector bundles with base space M, and with fibre over a point p of M consisting of tensors in M at p. These are the bundles which are particularly interesting in differential geometry. We now introduce them.

Let M be an $n-$dimensional manifold. Denote by B the set consisting of all pairs (p, ξ^a), where p is a point of M and ξ^a is a contravariant vector in M at p. Define a mapping π from the set B to the manifold M as follows: $\pi(p, \xi^a) = p$. For each point p of M, introduce the obvious $n-$dimensional vector space structure on $\pi^{-1}[p]$, i.e., $(p, \xi^a) + c(p, \eta^a) = (p, \xi^a + c\eta^a)$. Our plan is to make (B, M, π) into a vector bundle. What must be done is to introduce a manifold structure on B, show that π is then a smooth mapping, and, finally, show that B is, locally, a product. These three things are all done at once. Let (U, φ) be a chart on M, so each point of U is labeled by n coordinates, $x^1, \ldots, x^n$. A contravariant vector in M at this point can be labeled by its n components with respect to this chart, $\tilde{\xi}^1, \ldots, \tilde{\xi}^n$. Thus, each point of the subset $\pi^{-1}[U]$ of B is labeled by $2n$ numbers, $(x^1, \ldots, x^n, \tilde{\xi}^1 \ldots, \tilde{\xi}^n)$. This will be a chart on B. These charts are clearly compatible, so we are led to a manifold structure on B. Furthermore, since π, applied to the point of B with coordinates $(x^1, \ldots, x^n, \tilde{\xi}^1, \ldots, \tilde{\xi}^n)$, is the point of M with coordinates $(x^1, \ldots, x^n)$, π is certainly smooth. Finally, it is also clear that B is locally a product. Let V^n be labeled by $n-$tuples, $(\tilde{\xi}^1, \ldots \tilde{\xi}^n)$, and let ψ, applied to the point of $U \times V^n$ represented by $(p; \tilde{\xi}^1 \ldots, \tilde{\xi}^n)$. That this ψ has all the required properties is, once again, clear. In words, a chart on M, since contravariant vectors can then be expressed in terms of components, induces a natural chart on B.

Thus, any manifold M defines a vector bundle in which the base space is M, the bundle space is the collection of contravariant vectors at points of M, the fibre over a point p of M is the vector space of contravariant vectors at p, and the projection takes a contravariant vector at a point of M to the point at which that vector is. This vector bundle is called the *tangent bundle* of M, TM. (Note that the bundle structure is just right here. It is meaningless to say that a contravariant vector at one point of M is "the same" as a contravariant vector at another point of M. The notation of a vector bundle, conveniently, also does not allow such a comparison.)

Similarly, the vector bundle with base space M, fibre over the point p of M the $n-$dimensional vector space of covariant vectors at p, bundle space the collection of all covariant vectors at points of M, etc. is called the *cotangent bundle* of M, CM.

More generally, consider tensors with a given index structure, e.g., $\alpha^{ac}{}_{rq}{}^{d}$, at points of M. Using exactly the same argument as above, we have a vector bundle with base space M, and fibre over the point p of M tensors in M at p with precisely this index structure. (In this example, $k = n^5$.) The vector space structure in each fibre is, of course, the vector space structure on tensors at a fixed point p of M. These are called the *tensor bundles* of M. (They are, of course, still vector bundles in the sense of Sect. 16.)

Example 41. Let M be a manifold. Let B be the collection of pairs $(p, \alpha^{ac}{}_{rq}{}^{d})$, where $\alpha^{ac}{}_{rq}{}^{d}$ is a tensor at p satisfying $\alpha^{bc}{}_{nq}{}^{d} = \alpha^{a[c}{}_{nq}{}^{d]}$ (or any other linear condition on α). Then, as above, we obtain a vector bundle with base space M.

Example 42. Let M be a manifold. Let B be the collection of triples, $(p, \beta^a{}_b, \gamma^{de}{}_f)$, where $\beta^a{}_b$, and $\gamma^{de}{}_f$ are tensors at p. Set $\pi(p, \beta^a{}_b, \gamma^{de}{}_f) = p$. Define addition in the fibres by $(p, \beta^a{}_b, \gamma^{de}{}_f) + c(p, \tau^a{}_b, \mu^{de}{}_f) = (p, \beta^a{}_b + c\tau^a{}_b, \gamma^{de}{}_f + c\mu^{de}{}_f)$.

Then, as above we have a vector bundle with base space M.

Exercise 102. Why is there no vector bundle with "derivative operators" replacing tensors?. Why is there no vector bundle with "metrics" replacing tensor?

Exercise 103. Prove that all the tensor bundles of R^n are products.

Exercise 104. Using the fact that every vector field on S^2 vanishes at some point, prove that the tangent bundle of S^2 is not a product.
Exercise 105. Suppose we have a diffeomorphism between M and M'. Find a diffeomorphism between the bundle spaces of their cotangent bundles.
Exercise 106. Find a natural diffeomorphism between the bundle space of the tangent bundle of the product of two manifold and the product of the bundle spaces of the tangent bundles of the two manifolds.
Exercise 107. Why is R^7 not the bundle space of the cotangent bundle of any manifold?
Exercise 108. Show that a metric on M defines a natural diffeomorphism between the bundle space of the tangent bundle of M and the bundle space of the cotangent bundle of M. (Raising and lowering.)
Example 43. In classical mechanics, the configuration space of a system is a manifold M. The Lagrangian is a function on the tangent bundle of M. The cotangent bundle of M is called phase space, and the Hamiltonian is a function on the cotangent bundle.
Example 44. Let M be a manifold. There is a natural covariant vector field on the cotangent bundle of M. Let P be a point of CM, so P consists of a pair (p, μ_a), where μ_a is a covariant vector in M at the point $p = \pi(P)$ of M. A covariant vector in CT at P assigns, to each contravariant vector in CT at P, a real number. Let Λ be a contravariant vector in CT at P, and consider the number $\mu_a(\lambda^a)$ where $\lambda^a = \vec{\pi}_P \Lambda$ is a contravariant vector in M at p. This assignment defines a covariant vector at each point P of CM, and hence a covariant vector field (Exercise 109. Prove smooth.) on CM. In classical mechanics, the exterior derivative of this covariant vector field is the symplectic structure on phase space.

Let M be a manifold, and let B be one of the tensor bundles over M. A mapping which takes each point p of M to a point of the fibre over p certainly assigns a tensor of the appropriate type to each point of M, and conversely. In other words, there is a natural, one-to-one correspondence between (not necessarily smooth) tensor fields on M and (not necessarily smooth) mappings $\Delta : M \to B$ which satisfy $\pi \cdot \Delta =$ the identity diffeomorphism on M. We prove:
Theorem 15. The tensor field is smooth if and only if the mapping $\Delta : M \to B$ is smooth.

Proof: We do the case for the tangent bundle, all others being identical. Let (U, φ) be a chart on M. Then $\Delta : M \to B$ is represented by n functions of n variables: $\Delta(x^1, \ldots, x^n) = (x^1, \ldots, x^n, \tilde{\xi}^1(x), \ldots, \tilde{\xi}^n(x))$. Clearly, Δ is smooth if and only if, for every such chart, the functions $\tilde{\xi}^1(x), \ldots, \tilde{\xi}^n(x)$ are C^∞. But a contravariant vector field on M is smooth if and only if its components in every chart are C^∞ functions of the coordinates.

Let (B, M, π) be a vector bundle. A smooth mapping $\Delta : M \to B$ such that $\pi \cdot \Delta =$ identity on M is called a smooth *cross section*. Thus, for the tensor bundles, the cross sections are precisely the smooth tensor fields. We could have just as well have defined a smooth tensor field as one obtained as a smooth cross section, so smoothness of tensor fields would be obtained from smoothness of maps.
Exercise 110. prove that every vector bundle possesses a cross section.

The fibre over p in the tangent bundle (resp. cotangent bundle) is called the *tangent space* (resp. *cotangent space* of p).

18. Smooth Mappings: Action on Tensor Bundles

Let $\psi : M \to M'$ be smooth. Recall, from Sect. 15, that this ψ induces a mapping $\vec{\psi}_p$ from contravariant tensor in M at p to contravariant tensors in M' at $p' = \psi(p)$, a mapping $\underset{\leftarrow}{\psi}_p$ from covariant tensors in M' at p' to covariant tensors in M at p, and a mapping $\underset{\leftarrow}{\psi}$ from covariant tensor fields on M' to covariant tensor fields on M'. This discrimination against contravariant tensors is corrected, to a certain extent, when we ask what mappings ψ induces on the corresponding tensor bundles. We now ask this.

Let $\psi : M \to M'$ be smooth. Then, for each point p of M, $\underset{\leftarrow}{\psi}_p$ is a linear mapping from the cotangent space of $p' = \psi(p)$ to the cotangent space of p. Does this $\underset{\leftarrow}{\psi}_p$ induce a mapping from the cotangent bundle of M', CM', to CM? No. For one thing, certain points of M' may not even be in the range of ψ, so the fibres over such points would not have anywhere to be mapped to. Furthermore, two distinct points, p and q, of M might be mapped, by ψ, to the same point of M', whence there would be ambiguity as to where the cotangent space of this point of M' should be mapped to.

Of course, it works the other way around. Let p be a point of M, ξ^a a vector in M at p, so (p, ξ^a) is a point of TM. Then $(\psi(p), \vec{\psi}_p(\xi^a))$ is a point of TM', for $\vec{\psi}_p\ \xi^a$ is a contravariant vector in M' at the point $\psi(p)$ of M'. Hence, ψ induces a mapping $\overset{\rightharpoonup}{\psi} : TM \to TM'$. We prove that this mapping is smooth. Let (U, φ) be a chart on M, and (U', φ') a chart on M'. Then $x^1, \ldots, x^n$ are coordinates in U, while $x'^1, \ldots, x'^{n'}$ are coordinates in U'. In terms of these charts, the action of ψ is represented by n', C^∞ functions of n variables, $x'^1(x^1, \ldots, x^n), \ldots, x'^{n'}(x^1, \ldots, x^n)$. A point of TM in $\pi^{-1}[U]$ is represented by a $2n$-tuple of numbers $(x^1, \ldots, x^n, \tilde{\xi}^1, \ldots, \tilde{\xi}^n)$, and similarly for M'. In terms of these charts on TM and TM', the action of $\overset{\rightharpoonup}{\psi}$ is as follows: $\overset{\rightharpoonup}{\psi}(x^1, \ldots, x^n, \tilde{\xi}^1, \ldots, \tilde{\xi}^n) = (x'^1(x), \ldots, x'^{n'}(x), \Sigma_i \tilde{\xi}^i \frac{\partial x'^j}{\partial x^i}|_x \ldots, \Sigma_i \tilde{\xi}^i \frac{\partial x'^{n'}}{\partial x^i}|_x)$. That is, we have $2n'$ functions of $2n$ variable. Since these functions are C^∞, the mapping $\overset{\rightharpoonup}{\psi}$ is smooth.

The general situation is now clear. $\psi : M \to M'$ induces a smooth map, $\overset{\rightharpoonup}{\psi}$, from a tensor bundle over M to a tensor bundle over M', provided the tensor bundle over M consists of tensors all of whose indices are contravariant. No natural mappings are induced on the mixed bundles or covariant tensor bundles.

Finally, we note that $\psi \cdot \pi = \pi' \cdot \overset{\rightharpoonup}{\psi}$, and equation which states that the fibre over a point p of M is mapped, by $\overset{\rightharpoonup}{\psi}$ to the fibre over $\psi(p)$ of M'.

The second half of the table on p. 44 can now be written in a more symmetrical form:

$\psi : M \to M'$	contravariant tensor fields	covariant tensor fields	mixed tensor fields	contravariant tensor bundles	covariant tensor bundles	mixed tensor bundles
From M to M'	—	—	—	$\overset{\rightharpoonup}{\psi}$	—	—
From M' to M	—	$\underset{\leftarrow}{\psi}$	—	—	—	—

Of course, if $\psi : M \to M'$ is a diffeomorphism, then ψ induces a diffeomorphism between all the tensor bundles of the same index type, this diffeomorphism preserving all structure (i.e., their fibres, the vector-space structure, etc.).

Exercise111. Let $\psi : M \to M'$ and $\Lambda : M' \to M''$ be smooth. Prove that $\overset{\rightsquigarrow}{\Lambda} \cdot \overset{\rightsquigarrow}{\psi} = (\Lambda \overset{\rightsquigarrow}{\cdot} \psi)$.

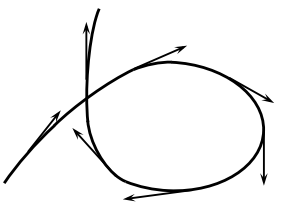
Exercise 112. Show that a metric on M induces a natural function $(g_{ab}\xi^a\xi^b)$ on TM. Prove that this function is smooth.

Exercise 113. In sect. 15, we argued that scalar fields should be regarded as covariant tensor fields of rank zero, for $\underset{\leftarrow}{\psi}$ acts on them. Introduce the vector bundle for which scalar fields are the cross sections, and show that $\overset{\rightsquigarrow}{\psi}$ acts on this bundle. Hence, scalar fields should also be regarded as contravariant tensor fields of rank zero.

Exercise 114. Let ψ be smooth from M to M'. Find an example in which ψ is onto but $\overset{\rightsquigarrow}{\psi}$ is not. In which ψ is one-to-one, but $\overset{\rightsquigarrow}{\psi}$ is not.

19. Curves

Let M be a manifold. By a (smooth) *curve* in M we understand a smooth mapping $\gamma : I \to M$ from some open interval $I = (a, b)$ a possibly $-\infty$ or b $+\infty$ of the real line to M. (Note that I, as an open subset of R^1, inherits a manifold structure.) If t is a real number in the interval I we write $\gamma(t)$ for the image of t under γ. Note that, by this definition, a choice of parameterization is part of a curve.



We define the tangent vector to a curve. Since the interval I is a subset of the reals R, the manifold I has an obvious natural chart on it. Denote by i^a the contravariant vector field on I whose component, with respect to this chart, is (1). Thus, if $f(t)$ (t in I) is a smooth function on I, then $i^a\ D(f)_a$ is the function df/dt on I. For each point t of I, $\vec{\gamma}_t(i^a)$ is, therefore, a contravariant vector in M at the point $\gamma(t)$ of M. This vector is called the *tangent vector* to the curve at t. Of course, the curve may "cross itself", i.e., we may have $\gamma(t_1) = \gamma(t_2)$ for $t_1 \neq t_2$. The tangent vectors at t_1 and t_2 may be different for such a crossing point, i.e., we may have $\vec{\gamma}_{t_1}\ (i^a) = \vec{\gamma}_{t_1}\ (i^a)$.

Example 45. In terms of a chart, a curve is represented by n functions of one variable, $x^1(t), \ldots, x^n(t)$. The tangent vector to this curve at t_0 is the vector at the point with coordinates $x^1(t_0), \ldots, x^n(t_o)$ with components $(dx^1/dt|_{t_0}, \ldots, dx^n/dt|_{t_0})$.

Example 46. Let p be a point of the manifold M, and let $\gamma : I \to M$ be the constant curve, $\gamma(t) = p$, which just remains at p. We show that its tangent vector is zero. If h is any smooth function on M, $\underset{\leftarrow}{\gamma}, h$ is a constant function on I, namely the one with value $h(p)$. So, $i^a D(\underset{\leftarrow}{\gamma} h)_a = 0$. So, $\vec{\gamma}_t\ i^a = 0$.

Example 47. Let f be a smooth function on M, and consider $f : M \to R$ as a smooth mapping of manifolds. Denote by i_a the covariant vector field on the manifold R whose component, in the natural chart, is (1). Then $\underset{\leftarrow}{f}(i^a) = D(f)_a$.

We are now in a position to give an intuitive geometrical interpretation of a contravariant vector at a point of a manifold M as an "infinitesimal displacement". Let p, be a point of M. In order to discuss infinitesimal displacements from p, we need a family of points of M which approach p. A curve does the job.

So, let $\gamma : I \to M$ be a curve, say, with $\gamma(t_0) = p$. Let ξ^a be the tangent vector to γ at t_0. Then, one should interpret "$\xi^a dt$ as the infinitesimal displacement vector from the point $p = \gamma(t_0)$ of M to the point $\gamma(t_0 + dt)$ of M." One justification for this interpretation would be the following. Set $\tilde{f} = \underset{\leftarrow}{\gamma}\, f$, where f is any smooth function on M. Thus, $\tilde{f}(t)$ is just a smooth function on I. Now,

$$\begin{aligned}\left.\frac{d}{df}\tilde{f}(t)\right|_{t_0} &= i^a D(\tilde{f})_a|_{t_0} = i^a D(\underset{\leftarrow}{\gamma} f)_a|_{t_0} = i^a \underset{\leftarrow}{\gamma} D(f)_a|_{t_0} \\ &= (\vec{\gamma}_p i^a) D(f)_a|_p = \xi^a D(f)_a|_p\end{aligned}$$

Whereas contravariant vectors have this interpretation as "infinitesimal displacements," no such interpretation is available for covariant vectors.

Example 48. The interpretation discussed above can also be obtained, more directly, from Example 45.

Thus, a metric, g_{ab}, on M associates, with each contravariant vector ξ^a at a point, a number, $R_{ab}\ \xi^a\, \xi^b$. The metric can therefore be thought of as describing a distance between infinitesimally nearby points (at least, if g_{ab} is positive-definite). Since a knowledge of $R_{ab}\ \xi^a\, \xi^b$ for every contravariant ξ^a at a point determines the metric g_{ab} uniquely,

a metric represents nothing more than the information of these distances between infinitesimally nearby points.

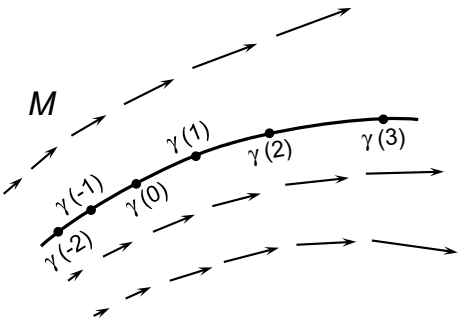
Example 49. Let g_{ab} be a positive-definite metric on M. We define the *length* of the curve $\gamma : I \to M$, where $I = (a, b)$. For each t in I, set $\lambda(t) = g_{ab}\, \xi^a\, \xi^b$, where ξ^a is the tangent vector to γ at t. Thus, $\lambda(t)$ is a function of one variable. Then (length of γ)= $\int_a^b [\lambda(t)]^{1/2} dt$. (Exercise 115. Verify that this length is independent of the parameterization. That is, let J be another open interval, and let $\mu : J \to I$ be a diffeomorphism. Then $\gamma \cdot \mu : \ J \to M$ is a curve in M. (It is practically the same curve. It hits the same points of M, but these are parameterized in a different way.) Verify that the length of $\gamma \cdot \mu$ is the same as that of γ.)

Finally, note that, if $\psi : \ M \to M'$ is a smooth mapping, and $\gamma : \ I \to M$ is a curve in M, then $psi \cdot \gamma : \ I \to M'$ is a curve in M'. Under smooth mapping, curves are sent the same way as contravariant tensors.

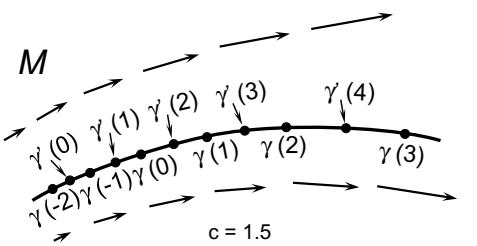
20. Integral Curves

Of course, a manifold M possesses a great number of curves. When a vector field is specified on M, then a certain collection of curves, defined by this vector field, are often of particular interest.

Let ξ^a be a contravariant vector field on M. A curve $\gamma : I \to M$ in M is said to be an *integral curve* of ξ^a if, for each point t of I, the tangent vector to $\gamma(t)$ at t (a contravariant vector in M at the point $\gamma(t)$ of M) coincides with the contravariant vector field ξ^a evaluated at $\gamma(t)$. Intuitively, an integral curve of a vector field "runs along in the direction of the vector field, moving quickly (i.e., covering a lot of M with each increment of t) where ξ^a is large, and slowly where ξ^a is small". If the interval I contains the origin 0 of R, then the point $\gamma(0)$ of M is called the *initial value* of the integral curve.



Let $\gamma : I \to M$ be an integral of ξ^a, with $I = (a, b)$. Let c be a real number, and set $I' = (a + c, b + c)$, and $\gamma'(t) = \gamma(t - c)$ for t in I'. Then $\gamma' : I' \to M$ is also an integral curve. (Proof: The diffeomorphism from I to I', which sends $t + c$ clearly takes the contravariant vector field i^a on I to i^a on I'.) If 0 is in both I and I', then the initial value of γ', $\gamma'(0)$. is the point $\gamma(-c)$ of M. Thus, all we have done is reparameterized the curve by adding a constant (namely c) to the perimeter. The consequence is to shift the initial value of $\gamma(t)$ to another point of the curve. It is clear that, given an integral curve, and a point p (in M) on that curve, one can, by reparameterizing in this way, obtain a new curve with initial value p. We shall call the operation described above *shifting the initial value*.



A fundamental property of integral curves is the following:

Theorem 16. Let ξ^a be a contravariant vector field on the manifold M, and let p be a point of M. Then there exists an integral curve of ξ^a, $\gamma : I \to M$, with initial value p, and with the following property: if $\gamma' : I' \to M$ is another integral curve of ξ^a, with initial value p, then $I' \subset I$, and for all t in I', $\gamma'(t) = \gamma(t)$

The proof of Theorem 16, although not very difficult, is rather technical and long. We omit it. A proof can be found in most textbooks on ordinary differential equations. It is immediate that the curve $\gamma : I \to M$ whose existence is guaranteed by Theorem 16 is unique, (Proof: If $\gamma : I' \to M$ were another, then we would have $I' \subset I$ and $I \subset I'$, so $I = I'$, and also $\gamma'(t) = \gamma(t)$ for t in I'. In other words, we would have the fact that γ' and γ are the same curve.) The unique curve obtained in Theorem 16 will be called the *maximal* integral curve (of ξ^a) with initial value p.

Theorem 16 is the basic existence and uniqueness theorem for ordinary differential equations. The following example illustrates this remark:

Example 50. Consider the pair of coupled ordinary differential equations

$$f' = \cos(\tau f) - g'\, e^g \tau^3 \qquad g'' = \frac{2f^2 g g' \tau}{f^2 + (g')^2} \tag{91}$$

for function $f\tau$) and $g(\tau)$, where a prime denote $d/d\tau$. Denote by M the four-dimensional manifold consists of R^4 with the (closed) region $(x^1) + (x^3)^2 = 0$ removed. We have a natural chart on M. Let ξ^a be the vector field on M whose components, with respect to this chart, are

$$\tilde{\xi}^1 = \cos(x^4 x^1) - x^3\, e^{(x^2)}(x^4)^3 \qquad \tilde{\xi}^3 = 2(x^1)^2\, x^2\, x^3\, x^4\, [(x^1)^2 + (x^3)^2]^{-1}$$
$$\tilde{\xi}^2 = x^3 \qquad \tilde{\xi}^4 = 1$$

This ξ^a is clearly smooth on M. Let p denote the point of M with coordinates $(a, b, c, 0)$. Then, by Theorem 16, there exists a unique maximal integral curve, $x^1(t), \ldots, x^4(t)$, of ξ^a, with initial value p. That is, by Example 45, we have

$$\frac{d}{dt}\, x^1 = \cos(x^4\, x^1) - x^3\, e^{(x^2)}\, (x^4)^3 \quad \frac{d}{dt}\, x^3 = 2(x^1)^2\, x^2\, x^3 x^4 \left[(x^1)^2 + (x^3)^2\right]^{-1} \tag{92}$$
$$\frac{d}{dt} x^2 = x^3 \qquad\qquad \frac{d}{dt} x^4 = 1$$

with

$$x^1(0) = a \qquad x^3(0) = c$$
$$x^2(0) = b \qquad x^4(0) = 0$$

Clearly, $x^4(t) = t$. Then, from (92), the functions $f(\tau) = x^1(\tau)$ and $g(\tau) = x^2(\tau)$, satisfy the differential equations (91). (Note that $x^3(\tau) = g'(\tau)$.) This solution has initial values $f(0) = a$, $g(0) = b$, and $g'(0) = c$. In other words, Theorem 16 ensures that, given arbitrary values of f, g, and g' at $\tau = 0$, (with $(f(0))^2 + (g'(0))^2 = 0$), the differential equation (91) has precisely one solution extended maximally in τ.

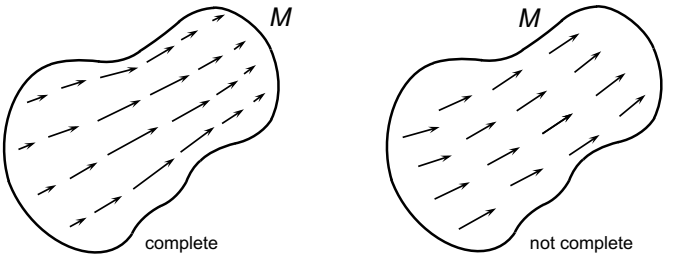
A number of properties of integral curves follow immediately from Theorem 16. Let $\gamma : I \to M$ be a maximal integral curve, with $I = (a, b)$ (a may be $\mp\infty$, or $b + \infty$, or both). Let d be a number in the interval (a, b). Then, as we remarked before, $\gamma' : J \to M$, defined by $\gamma'(t) = \gamma(t + d)$ where $J = (a - d,\ b - d)$, is an integral curve. In fact, γ' is maximal. (Proof: If the $t-$interval J of γ' could be enlarged, then, by shifting the initial value, we would have an enlargement of the $t-$interval I of γ. But this would contradict the assumed maximality of γ.) Thus, shifting of the initial value, applied to a maximal integral curve, results in a maximal integral curve.

Integral curves cannot cross. That is, if $\gamma : I \to M$ and $\gamma' : I' \to M$ are maximal integral curves, and if $\gamma(t_1) = \gamma'(t_2) = p$ for some t_1 and t_2, then γ and γ' are obtainable from each other by shifting the initial value. Proof: By shifting initial values, we can have p the initial value of γ and of γ'. By Theorem 16, these curves are then identical. Hence, the original curves differ from each other only by shifting of the initial value.

As a final illustration of the use of Theorem 16, we remark that, if p is a point of M at which ξ^a vanishes, and if $\gamma : I \to M$ is an integral curve passing through p (i.e., if $\gamma(t) = p$ for some t), then γ is the constant curve: $\gamma(t) = p$ for all t. Proof: By shifting the initial value, we might as well take p as the initial value of γ. Let $\tilde{\gamma} : R \to M$ be the constant curve, $\tilde{\gamma}(t) = p$ for all t. Then, by Example 46, $\tilde{\gamma}$ is an integral curve of ξ^a. Since the interval is the whole real line, this integral curve must be maximal. Hence, by Theorem 16, γ is the constant curve remaining at p.

Numerous other – intuitively clear – properties of integral curves follow directly and similarly from Theorem 16.

We next discuss the dependence of the maximal integral curves on the initial value. For this purpose, it is convenient to introduce the following definition. A contravariant vector field ξ^a on M is said to be *compete* if every maximal integral curve has $I = (-\infty,\ +\infty)$. Intuitively, a complete vector field has the property that ξ^a becomes "small" as one approaches the "edge" of M, so only "in the limit $t \to \infty$ or $t \to -\infty$ do integral curves approach the edge". (Note: It is more conventional to define completeness



of a vector field by requiring only that the intervals have the form $(a, +\infty)$, where a may be finite or $-\infty$,)

Exercise 116. Let ξ^a be a complete on M, and let C be a closed subset of M. Prove that ξ^a is complete on $M - C$ and if and only if every integral curve with initial value in $M - C$ remains in $M - C$.

Exercise 117. Let ξ^a be complete on M, and let p be a point of M. Prove that ξ^a is complete on $M - p$ if and only if ξ^a vanishes at p.

Exercise 118. Find an example of two complete vector fields whose sum is not complete; of two vector fields neither of which is complete, but whose sum is complete.

Exercise 119. Prove that ξ^a is complete if and only if $-\xi^a$ is complete.

Let ξ^a be a complete vector field on M. We define a mapping $\Gamma : R \times M \to M$ as follows. If t is a number, p a point of M, $\Gamma(t, p) = \gamma(t)$, where γ is the maximal integral curve with initial value p. (Exercise 120. Why do we need completeness for this?) The second fundamental property of integral curves is the following:

Theorem 17. Let ξ^a be a complete contravariant vector field on M. Then $\Gamma : R \times M \to M$ is smooth.

Again, we omit the proof. See most textbooks on ordinary differential equations. In terms of differential equations, Theorem 17, states that the solution of s differential equation depends smoothly on the initial value.

Write $\Gamma_t(p)$ for $\Gamma(t, p)$. Then, fixed t, $\Gamma_t : M \to M$ is a mapping from M to M. Intuitively, the action of Γ_t is as follows: "move each point p of M a parameter-distance t along the integral curve through p". Thus, points of M where ξ^a is "large" are "moved a great deal" by Γ_t, points where ξ^a is small are moved less far, and points of M at which ξ^a vanishes are left invariant by Γ_t.

Exercise 121. Prove that, if ξ^a vanishes at p, then $\Gamma_t(p) = p$ for all t.

We prove that Γ_t is smooth for each t. Fix t, and let Λ; $M \to R \times M$ be defined by $\Lambda(p) = (t, p)$. Then Λ is smooth. Furthermore, $\Gamma_t = \Gamma \cdot \Lambda$. By Theorem 17, Γ_t is smooth. Note, furthermore that, by shifting the origin, Γ_{-t} is the inverse of Γ_t (i.e., $\Gamma_{-t} \cdot \Gamma_t = \Gamma \cdot \Gamma_{-t} =$ identity mapping on M). Thus, Γ_t is one-to-one and onto. It's inverse, Γ_{-t}, is also smooth (for Γ_ω is smooth for all ω, in particular, for $\omega = -t$). Thus,

Theorem 18. For each number t, $\Gamma_t : M \to M$ is a diffeomorphism. Furthermore, $\Gamma_t \cdot \Gamma_\omega = \Gamma_{(t+\omega)}$.

These remarks further strengthen the interpretation of contravariant vectors as "infinitesimal displacements". The mapping Γ_t represents the corresponding "finite displacement" obtained by "integrating the infinitesimal displacement defined, at each point, by ξ^a". So, "Γ_{dt} represents the corresponding infinitesimal displacement".

Exercise 122. Consider the vector field in R^2 with components $\tilde{\xi}^1 = x^2, \tilde{\xi}^2 = -x^1$. Prove that this field is complete. Find an explicit formula for Γ_t. Check Theorem 17 and 18 explicitly.

Example 51. In classical mechanics, the vector field $\xi^a = F^{ab}\ D(H)_t$, in phase space, where F^{ab} is the inverse of the symplectic 2−form at each point (i.e., $F^{ab}\ F_{cb} = \delta^a{}_c$) and H is the Hamiltonian function, is called the Hamiltonian vector field. The integral curves of this (normally complete) vector field describe the evolution of the system in time.

Example 52. In fluid mechanics, the velocity field of the fluid is a complete vector field in Euclidean 3−space. Then Γ_t maps the locations of fluid elements at time zero to their

locations at time t. A vector field is sometimes called a flow.
Exercise 123. Find a nonzero vector field such that Γ_t = identity for some t.
Exercise 133. Find an example of a diffeomorphism from a manifold to itself which is not a Γ_t for any vector field or any t.
Exercise 134. Prove Theorems 16 and 17 in the case when M is one-dimensional.

21. The Lie derivative: Geometrical Interpretation

We have already discussed the Lie derivative from two different points of view: the algebraic approach (Sect. 7), and the concomitant approach (Sect. 11). The most intuitive, and often the most useful, approach is the one we now introduce.

Let ξ^a be a complete vector field on M. (Completeness merely serves to simplify the discussion here. It makes the range and domain of Γ_t be M. With slightly more effort, but with little increase in content, one could drop the completeness assumption in this section.) Let $\alpha^{\cdots}{}_{\cdots}$ be a tensor field on M. Now, for each number t, Γ_{-t} is a diffeomorphism on M. Thus, Γ_{-t} takes the tensor field $\alpha^{a\ldots c}{}_{b\ldots d}$ on M to some other tensor field, $\alpha^{a\ldots c}{}_{b\ldots d}(t)$, on M. That is to say, we have for each t, a tensor field $\alpha^{a\ldots c}{}_{b\ldots d}(t)$ on M. That is to say, we have, for each t, a tensor field $\alpha^{a\ldots c}{}_{b\ldots d}(t)$ on M. One says that $\alpha^{a\ldots c}{}_{b\ldots d}(t)$ results from $\alpha^{a\ldots c}{}_{b\ldots d}(0)$ (our original tensor field) by dragging along the vector field ξ^a.

We propose to prove that

$$\frac{d}{dt}\alpha^{a\ldots c}{}_{b\ldots d}(t)\Big|_{t=0} = \mathscr{Z}_\xi \alpha^{a\ldots c}{}_{b\ldots d} \tag{93}$$

where the left side has the following meaning. Fix a point p of M. Then, for each t, $\alpha^{a\ldots c}{}_{b\ldots d}(t)$ is a tensor at p. Then $\frac{d}{dt}\alpha^{a\ldots c}{}_{b\ldots d}(t)|_0$ is the tensor field whose value at p is

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\Big[\alpha^{a\ldots c}{}_{b\ldots d}(\Delta t) - \alpha^{a\ldots c}{}_{b\ldots d}(0)\Big]$$

Fix a point p of M, and let $\gamma: R \to M$ be the (maximal) integral curve with initial value p. We first establish (93) for a scalar field α. Let $\alpha(t) = \underset{\leftarrow}{\Gamma}_t\alpha$, and let $f(t)$ be the scalar field $\alpha(t)$ evaluated at p, so $f(t)$ is one real function of the variable. Then

$$\begin{aligned}\frac{d}{dt}f\Big|_0 &= i^a\, D(f)_a|_0 = i^a\, D(\underset{\leftarrow}{\Gamma}_t\,\alpha)_a|_0 = i^a \underset{\leftarrow}{\Gamma}_t\, D(\alpha)_a|_0\\ &= \Big((\vec{\Gamma}_t)_p i^a)\Big)\, D(\alpha)_a\Big|_p = \xi^a D(\alpha)_a\Big|_p = (Z)_\xi\alpha\Big|_p\end{aligned}$$

Since this equation holds for every point p, Eqn. (93) holds for scalar fields. Next, let $\omega_{a\ldots c}$ be a form. Then, since exterior differentiation is preserved by a diffeomorphism, we have, for every t,

$$\Gamma_t D(\omega)_{ma\ldots c} = D\,(\Gamma_t\omega)_{ma\ldots c} \tag{94}$$

Taking d/dt of this equation, and evaluating at $t = 0$,

$$\frac{d}{dt}\,D\Big(\omega(t)\Big)_{ma\ldots c} = D\Big(\frac{d}{dt}\omega(t)\Big)_{ma\ldots c} \tag{95}$$

Let α be a scalar vector. Then

$$\frac{d}{dt}D(\alpha)_a\Big|_0 = D(\frac{d}{dt}\alpha)_a\Big|_0 = D(\mathscr{Z}_\xi\,\alpha)_a = \mathscr{Z}_\xi\, D(\alpha)_a$$

where we have used (95) and the fact that the Lie derivative and exterior derivative commute. Thus, (93) holds for a gradient. If α and β are scalar field, therefore,

$$\begin{aligned}\frac{d}{dt}\Big(\beta(\alpha)_a\Big)\Big|_0 &= \Big(\frac{d}{dt}\beta\Big)\Big|_0 D(\alpha)_a + \beta\frac{d}{dt}D(\alpha)_a\Big|_0 = (\mathscr{Z}_\xi\beta)\, D(\alpha)_a + \beta\mathscr{Z}_\xi D(\alpha)_a\\ &= \mathscr{Z}_\xi\Big(\beta D(\alpha)_a\Big)\end{aligned}$$

But, since every covariant vector field can, locally, be expressed as a sum of products of scalar fields and gradients, Eqn. (93) holds for arbitrary covariant vector fields. Next, let η^a be a contravariant vector field, μ_a a covariant vector field. Then

$$\mu_a \frac{d}{dt}\eta^a\Big|_0 - \frac{d}{dt}(\mu_a\eta^a)\Big|_0 - \eta^a \frac{d}{dt}\mu_a\Big|_0 = \mathscr{L}_\xi(\mu_a\eta^a) - \eta^a \mathscr{L}_\xi\mu_a$$
$$= \mu_a \mathscr{L}_\xi \eta^a$$

Since μ_a is arbitrary, (93) holds for $\alpha^{a\ldots c}{}_{b\ldots d}$ a contravariant vector field. Finally, for an arbitrary tensor field, we proceed as usual:

$$\mu_a \ldots \nu_c\, \eta^b \ldots \tau^d \left.\frac{d}{dt}\alpha^{a\ldots c}{}_{b\ldots d}\right|_0 = \frac{d}{dt}[\alpha^{a\ldots c}{}_{b\ldots d}\mu_a \ldots \nu_c\eta^b \ldots \tau^d]|_0$$
$$- \alpha^{a\ldots c}{}_{b\ldots d}\left[(\frac{d}{dt}\mu_a)\ldots \nu_c\eta^b \ldots \tau^d + \ldots + \mu_a \ldots n u_c \eta^b \ldots \frac{d}{dt}\tau^d\right]\Big|_0$$
$$= \mu_a \ldots \nu_c \eta^b \ldots \tau^d \mathscr{L}_\xi \alpha^{a\ldots c}{}_{b\ldots d}$$

nothing that, since $\mu_a, \ldots, \nu_c, \eta^b, \ldots, \tau^d$ are arbitrary, (93) is proven.

Thus, ξ^a defines a one-parameter family of motions on M, each of which is a diffeomorphism. Each diffeomorphism carries tensor fields on M to fields on M. The "rate of change" of a tensor field under these motions is the Lie derivative of that field in the ξ^a-direction.

The fact that Lie derivatives habitually commute with other operations (e.g., exterior derivatives, contraction) should now be clear. Similarly, the Leibnitz rule for Lie derivatives, etc. is immediate from (93).

Exercise135. Let ξ^a be a vector field, and ∇_a derivative operator on M. Then, for each t Γ_{-t} takes ∇_a to another derivative operator, $\nabla_a(t)$, on M. These two operators are related by some tensor field, $\gamma^m{}_{ab}(t)$, on M. Prove that

$$\frac{d}{dt}\gamma^m{}_{ab}\Big|_0 = -\nabla_a\nabla_b\xi^m + R_{sab}{}^m\xi^s$$

The remarks surrounding Eqn. (66) now make sense.

Exercise136. Do Exercise 82 in one sentence.

Example 53. In general relativity, because of the interpretation above, Killing vectors represent symmetries in space-time.

Exercise 137. Let $\Gamma: R \times M \to M$ be smooth, and suppose that, for each point p of M, $\Gamma(t, \Gamma(w, p)) = \Gamma(t + w, p)$. Find a contravariant vector field on M such that it generates Γ (in the sense of Theorem 17).

Exercise 138. Prove Eqn. (31) directly from (93).

Exercise 139. In Exercise 122, let g_{ab} be the metric with components $g_{11} = g_{22} = 1, g_{12} = g_{21} = 0$. Prove that $\mathscr{L}_\xi g_{ab} = 0$ in two ways, once directly (e.g., using concomitants), and once using (93).

22. Lie Groups

As an example and application of some of the preceding material, we obtain a few properties of Lie groups.

Roughly speaking, a Lie group is both a group and a manifold, where these two structures interact in a natural way. Let G be a set. Suppose that we are given a group structure on G, i.e., suppose we are given a mapping $\mathscr{C} : G \times G \to G$ (composition) and a mapping $\mathscr{J} : G \to G$ (inversion), subject to the usual axioms for a group. One normally writes $\alpha\beta$ instead of $\mathscr{C}(\alpha,\beta)$and $\alpha-1$ instead of $\mathscr{J}(\alpha)$. Suppose further that we are given a manifold structure on G, i.e., suppose we are given a collection of charts on G satisfying $M1 - M4$. We now have both structures on G: we have yet to require that they interact properly with each other. The most natural "interaction" available is the requirement that the mapping $\mathscr{C} : G\times G \to G$ and $\mathscr{J} : G \to G$ (which define the group structure) are smooth (as mappings of manifolds). Thus, a *Lie group* is a set G with both a group and a manifold structure, such that $\mathscr{C}$ and $\mathscr{J}$ are smooth.

Example 54. The Lorentz group, the Poincaré group, $SU(3)$ are all Lie groups.

Example 55. Let V be a $k-$dimensional vector space. Considering V as a manifold R^k (by choosing a basis, so each element of V is defined by its components, and so represents an element of R^k), and as an abelian group (under addition of vectors), V^k becomes a Lie group.

Fix an element α of the Lie group G, and let $\Delta_\alpha : G \to G$ be defined by $\Delta_\alpha(\beta) = \alpha\beta$, for any β in G. Then, evidently, we have $\Delta_\alpha \cdot \Delta_\gamma = \Delta_{\alpha\gamma}$, and Δ_e = identity mapping from G to G, where e is the identity element of the group G. In particular, for each α in G, $\Delta_\alpha \cdot \Delta_{\alpha^{-1}} = \Delta_{\alpha^{-1}} \cdot \Delta_\alpha$ = identity mapping from G to G. Thus, each Δ_α is a one-to-one onto mapping from G to G. We show that Δ_α, for fixed α, is a diffeomorphism. Note that the mapping $\lambda : G \to G \times G$ which sends β to (α,β) is smooth (injection of one factor into the product). But $\Delta_\alpha = \mathscr{C} \cdot \lambda$, and so Δ_α, as a composition of smooth maps, is smooth. Thus, since Δ_α is smooth, one-to-one, and onto, with smooth inverse, Δ_α is a diffeomorphism. Note, also that $\Delta_\alpha(e) = \alpha$.

A tensor field $\mu^{a\ldots c}{}_{b\ldots d}$ on G will be said to be *left invariant* if, for each α in G, the diffeomorphism Δ_α takes the tensor field $\mu^{a\ldots c}{}_{b\ldots d}$ to itself. How does one obtain left invariant tensor fields on a Lie group? Let $\mu^{a\ldots c}{}_{b\ldots d}$ be any tensor in G at the point e (the identity). For each α in G, Δ_α is a diffeomorphism on G which takes e to α. Hence, Δ_α takes the tensor $\mu^{a\ldots c}{}_{b\ldots d}$ at e to some tensor at α. Repeating this for each point α of G we obtain a tensor at each point of G. That is, we obtain a tensor field on G. In fact, this tensor field is left invariant. This is clear, for, if α and β are two points of G, then Δ_β takes the tensor at e to the tensor at β, while $\Delta_{\alpha\beta} = \Delta_\alpha \cdot \Delta_{beta}$ takes the tensor at e to the tensor at $\alpha\beta$. Hence, Δ_α takes the tensor at β to the tensor at $\alpha\beta$. Since β is arbitrary, the e tensor field remains invariant under Δ_α. Since α is arbitrary, the tensor field is left invariant. We have:

Theorem 19. Given a tensor at e, there is precisely one left invariant tensor field on G which coincides with the given tensor at e.

It follows immediately that all the tensor bundles of a Lie group are product bundles. Thus, for example, we can conclude that S^2 cannot be the manifold of any Lie group (Exercise 104).

We define the Lie algebra of a Lie group. Intuitively, the Lie algebra is obtained by "commuting elements which differ infinitesimally from the identity". This remark suggests that we consider contravariant vectors at the identity e of G. Let ξ^a and η^a be contravariant vectors in G at e. Then, by Theorem 19, we have left invariant vector fields ξ^a and η^a on G. Clearly, $\mathscr{Z}_\xi\eta^a = [\xi,\eta]^a$ is also left invariant. Evaluating at e, we obtain a contravariant vector in G at e. Thus, with two contravariant vectors at e, we associate a

third, where this association is clearly linear. Thus, we have a tensor $C^m{}_{ab}$ at e such that $C^m{}_{ab}\xi^a\eta^b$ is the contravariant vector at e defined, as above, from ξ^a and η^a. This $C^m{}_{ab}$ is called the *structure constant tensor* of G. Equations (24) and (25) imply

$$C^m{}_{ab} = C^m{}_{[ab]} \tag{96}$$

$$C^m{}_{n[a}C^n{}_{bc]} = 0 \tag{97}$$

respectively. Eqn. (96) is obvious. For (97), note that $\Big[[\eta, \lambda],\, \xi\Big]$ is represented by $C^m{}_a\xi^a$, $C^n{}_{bc}\eta^b\lambda^c$.

A (finite-dimensional) *Lie algebra* is a finite-dimensional vector space with a tensor $C^m{}_{ab}$ over that space, satisfying (96) and (97). Thus, every Lie group defines a Lie algebra.

Set $g_{ab} = -C^m{}_{na}, C^n{}_{mb}$, a tensor in G at e. By Theorem 19, we obtain an, obviously symmetric, tensor field g_{ab} on G. This is called the *invariant metric* of the Lie group G. (It many not be a metric as we have defined this term, for g_{ab} may not be invertible.)

Example 56. $SU(2)$ is the Lie group of all complex, unitary, 2×2 matrices with unit determinant. It is underlying manifold is S^3. Hence, the tangent bundle of S^3 is a product. Choose a contravariant vector at e in $SU(2)$, and, by Theorem 19, obtain a nowhere vanishing vector field on S^3. This is called the Hopf fibration of S^3.

Exercise 140. Let G and G' be Lie groups. Show that $G \times G'$, considered as a product of manifolds and a direct sum of groups, is a Lie group.

Exercise 141. Prove that the structure constant tensor of a commutative group vanishes.

Exercise 142. Find all two-dimensional Lie algebras.

Exercise 143. Show that, in the definition of a Lie group, it suffices to require that the mapping from $G \times G$ to G which sends (α, β) to $\alpha\beta^{-1}$ be smooth.

Exercise 144. Find a Lie group whose manifold is the torus; the cylinder.

23. Groups of Motions

We have seen in Sect 20 that a contravariant vector field on a manifold M yields a collection Γ_t (t a number) of diffeomorphisms on M, where these diffeomorphisms satisfy $\Gamma_t \cdot \Gamma_w = \Gamma_{t+w}$ and $\Gamma_0 =$ identity on M. In other words, the additive group of real numbers is realized as a group of diffeomorphisms on M. (More generally, the collection of all diffeomorphisms on a manifold M forms a group, with composition as the group operation. Above, we obtain a subgroup isomorphic with the additive group of reals.) One can consider more general groups realized as diffeomorphisms on a manifold. We briefly describe this type of situation.

Let G be a lie group, and M a manifold. Consider a smooth mapping $\Gamma: G \times M \to M$. Writing $\Gamma_\alpha(p)$ for $\Gamma(\alpha, p)$ (so, for each α in G, Γ_α is a smooth mapping from M to M), we require that composition of these mappings reflect the group operation in G, i.e., we require that

$$\Gamma_\alpha \cdot \Gamma_\beta = \Gamma_{\alpha\beta} \qquad \Gamma_c = \text{identity} \tag{98}$$

A *group of motions* on M consists of a Lie group G and a smooth mapping $\Gamma: G \times M \to M$ satisfying (98). By the same argument used in this situation in Sects. 20 and 22, we see that each is a diffeomorphism on M.

Example 57. Every Lie group defines a group of motions on itself.

Example 58. A group of motions on a manifold M defines a group of motions on TM.

Example 59. The Poincaré group is a group of motions on Minkowski space (a 4−manifold).

Example 60. A complete contravariant vector field on M defines, as in Sect 20, a group of motions on M, where G is the additive group of real numbers.

Example 61. The Lie group of nonsingular linear mapping on a finite-dimensional vector space is a group of motions on that vector space (considered as a manifold).

A group of motions is said to be *effective* if $\Gamma_\alpha =$ identity only when $\alpha = e$, and *transitive* if, for any two points p and q of M, there is an α in G such that $\Gamma_\alpha(p) = q$.

Let G be a group of motions on M. Intuitively, a contravariant vector in G at e represents "an element of G differing infinitesimally from e". Hence, it defines an "infinitesimal motion on M". That is, it takes "each point of M to a nearby point". Thus, we should have a contravariant vector at each point of M, i.e., a contravariant vector field on M. Thus, a contravariant vector in G at e should define a contravariant vector field on M. We now define this field.

Let ξ^a be a contravariant vector in G at e. Fix a point p of M. Let $\psi: G \to G \times M$ be defined by $\psi(\alpha) = (\alpha, p)$, Then ψ sends e to the point (e, p) of $G \times M$. Hence, ψ sends ξ^a to some contravariant vector in $G \times M$ at (e, p). But $\Gamma: G \times M \to M$ sends (e, p) to p. Hence, $(\Gamma \cdot \vec{\psi})_e \xi^a$ is a contravariant vector in M at p. Repeating for each p, we obtain a contravariant vector field on M. Thus, each ξ in G at p defines a field ξ'^a on M.

Let ξ^a and η^a be contravariant vector in G at e. We shall show that

$$[\xi, \eta]' = \mathscr{L}_{\xi'} \eta' \tag{99}$$

That is if the bracket of ξ^a and η^a (in G) is taken to a field on M, the result is the same as the Lie derivative of η'^a in the ξ'^a−direction (i.e., Lie derivatives of fields in M). Take ξ^a as a vector field on G, and let γ_t (t a number) be the corresponding 1−parameter group of motion on G. Then, evidently, $\Gamma_{\gamma_{t(a)}}$ is a 1−parameter group of motions on M. Consider η^a as a contravariant vector field on G. Then, for each t, γ_t takes η^a to some other contravariant vector field on G. Let κ^a be that field evaluated at e. Then κ'^a is the vector field on M obtained by acting on η'^a with the diffeomorphism $\Gamma_{\gamma_{t(t)}}(e)$ on M. Taking d/dt of this equality at $t = 0$, and using (93), Eqn, (99) follows.

To summarize, a group of motions on M (where G is k−dimensional) defines a k−dimensional vector space of contravariant vector fields on M. This vector space is

naturally isomorphic to the tangent space to G at e. Lie brackets in G correspond to Lie derivatives of these vector fields) in M. Thus, if a Lie group is realized as a group of motions on M, its Lie algebra is realized as a Lie algebra of contravariant vector fields on M. Note that Sect. 20 is the special case when G is one-dimensional.

Example 62. For the Poincaré group as a group of motions on Minkowski space, the corresponding vector fields are the Killing vectors on Minkowski space.

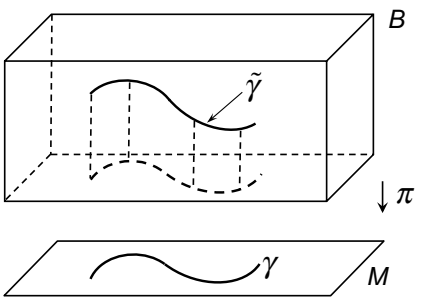
Example 63. G is effective if and only if the mapping from the tangent space to G at e to vector fields on M is one-to-one. For G to be transitive, it is necessary that $\dim G \geq \dim M$.

Exercise 145. For the translations and rotations on the plane (a group of motions), find the corresponding vector fields and their Lie derivatives explicitly.

24. Dragging Along

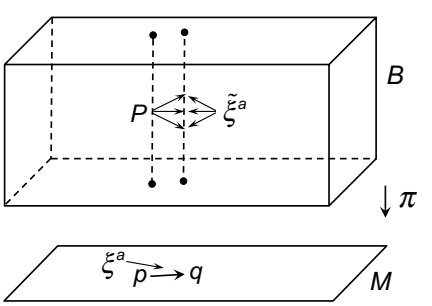
Let ξ^a be a contravariant vector field on the manifold M. Then, through each point of M, there passes an integral curve of ξ^a. Now suppose we are given a tensor at a point of M. We wish to transport that tensor along the integral curve of ξ^a (i.e., define a tensor at each point of the curve) by the requirement that the Lie derivative of the tensor in the ξ^a−direction be zero. From the concomitant expression (52), we see that this amounts to solving a certain ordinary differential equation. But we have a method, Theorem 16, for solving ordinary differential equations (or, at least, for asserting existence and uniqueness of solutions). Thus, we wish to arrange matters so that transporting the tensor amounts to finding integral curves of some vector field.

It is convenient to first discuss this problem in a somewhat more general setting. Let B be one of the tensor bundles over M. Let $\gamma : I \to M$ be a curve in M. Then a curve $\tilde{\gamma} : I \to B$ is said to be a *lifting* of γ if $\gamma = \pi \cdot \tilde{\gamma}$. Thus, if $\tilde{\gamma}$ is a lifting of γ, then, for each t in I, $\tilde{\gamma}(t)$ is a point of the fibre over $\gamma(t)$. So a lifting of γ amounts to selecting, for each t, some point in the fibre over $\gamma(t)$. In other words, a lifting of γ is essentially a choice, for each t, of a tensor in M, at $\gamma(t)$. Clearly, a lifting represents a specification of how to transport a tensor along γ.



Now let ξ^a be a contravariant vector field in M, and B one of the tensor bundles over M. In order to transport tensors along the integral curves of ξ^a, we must obtain a lifting of each integral curve of ξ^a. A contravariant vector field $\tilde{\xi}^a$ in B will be said to be a *lifting* of ξ^a if, for each point P of B, $\vec{\pi}_P\tilde{\xi}^a = \xi^a$. How unique is the lifting of a vector field? If $\tilde{\xi}^a$ is a lifting of ξ^a, we can add to $\tilde{\xi}^a$ at P any other vector λ^a with $\vec{\pi}_P\lambda^a = 0$. But, as we have seen, there is a k−dimensional vector space of such vectors at P (where k = dimension of the fibres): these are the vertical vectors which lie in the fibres. Geometrically, if ξ^a at p connects p to the "infinitesimally displaced point q", then $\tilde{\xi}^a$ at P (where P is in the fibre over p) must connect P with "any infinitesimally displaced point (of B) in the fibre over q." The choice of the point of the fibre over q is the same as the choice of a λ^a lying within the fibres.

Liftings of vector fields and curves are related by the – intuitively clear – statement:



Theorem 20. Let ξ^a be a contravariant vector field on M, B one of the tensor bundles over M, and $\tilde{\xi}^a$ a lifting of ξ^a. Then each integral curve of $\tilde{\xi}^a$ (in B) is a lifting of an integral curve of ξ^a (in M).

Proof: Let $\gamma : I \to B$ be an integral curve of $\tilde{\xi}^a$, and let $\pi \cdot \tilde{\gamma} : I \to M$ be the corresponding curve in M. We must show that $\pi \cdot \tilde{\gamma}$ is an integral curve of ξ^a. But this is immediate, for the tangent vector to $\pi \cdot \tilde{\gamma}$ at t is $(\vec{\pi \cdot \tilde{\gamma}})_t = \vec{\pi}_P \cdot \vec{\tilde{\gamma}}_t i^a = \vec{\pi}_P\tilde{\xi}^a = \xi^a$, where $P = \tilde{\gamma}(t)$, $p = \pi(P)$.

We now return to the original problem, that of obtaining a mode of transport using the Lie derivative. Let ξ^a be a contravariant vector field in M, B one of the tensor bundles over M. From the discussion above, the problem reduces to that of specifying a particular lifting of ξ^a. For each t, Γ_t is a diffeomorphism on M. Hence, Γ_t defines a diffeomorphism, $\tilde{\Gamma}_t$ on B. Of course, $\tilde{\Gamma}_t$ is fibre preserving: $\pi \cdot \tilde{\Gamma}_t = \Gamma_t \cdot \pi$. Furthermore, $\tilde{\Gamma}_t$ preserves the

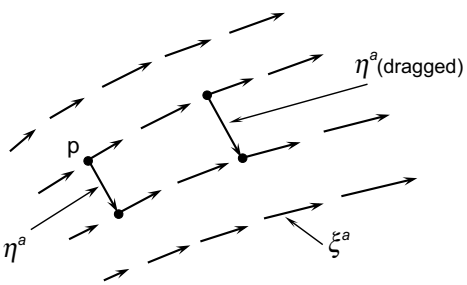
vector-space structure within each fibre. As in Sect. 23, the one-parameter group of motions $\tilde{\Gamma}_t$ on B defines a contravariant vector field $\tilde{\xi}^a$ in B. By the usual argument, $\tilde{\xi}^a$ is a lifting of ξ^a. Thus, there is a natural lifting of any contravariant vector field in M.

Let $(p, \alpha^{a\ldots c}{}_{b\ldots d})$ be a point of B, so $\alpha^{a\ldots c}{}_{b\ldots d}$ is a tensor in M at p. Then the integral curve of $\tilde{\xi}^a$ through $(p,\ \alpha^{a\ldots c}{}_{b\ldots d})$ is a lifting of the integral curve of ξ^a through p. Thus, the first integral curve defines, at each point of the second integral curve, a tensor in M at that point. In other words, given a tensor in M at p, we obtain a tensor at each point of the integral curve of ξ^a through p. These tensors are said to be the result of *dragging along* of $\alpha^{a\ldots c}{}_{b\ldots d}$.

Clearly, a tensor field on M is equal to the result of dragging along that tensor from any point of M if and only if the Lie derivative of that tensor field in the ξ^a-direction is zero. In this sense, dragging along represents "transport by the requirement that the Lie-derivative in the ξ^a-direction be zero".

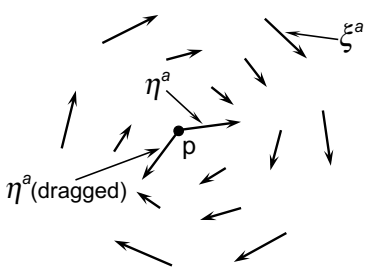
Example 64. Geometrically, dragging along of a contravariant vector field η^a can be interpreted as follows. Think of η^a as an infinitesimal displacement. Move each of these two nearby points a parameter-distance t along their respective integral curves of ξ^a. The resulting infinitesimal displacement is the result of dragging along of η^a. Thus, for the ξ^a vanishes at p, so the integral curve through p remains there. But dragging along of a contravariant vector at p results in a rotation of that vector, keeping it at p.

Exercise 146. Prove that the sum of the result of dragging along two tensors is the same as the result of dragging along their sum.



Exercise 147. Let ξ^a Be a contravariant vector field on M, and suppose that Γ_{t_0} = identity, so the integral curves of ξ^a are all closed loops. Prove that the result of dragging along a tensor around one of these closed loops is to leave the tensor invariant.

Exercise 148. Find an example of a ξ^a having an integral curve which is a closed loop, such that the result of dragging along a tensor around this closed loop does not leave the tensor invariant.



25. Derivative Operators: Interpretation in the Tensor Bundles

In Sect. 24. we saw that there is a natural lifting of any contravariant vector field $\hat{\xi}^a$ on a manifold to a contravariant vector field $\tilde{\xi}^a$ on each of the tensor bundles of M. Now suppose we have a derivative operator ∇_a on M. What additional structure appears on the tensor bundles? This question is the subject of the present section.

It is convenient to first introduce some notation. Let M be a manifold, B the tensor bundle of tensors with index structure $\alpha^{a\ldots c}{}_{b\ldots d}$. Let $K_a{}^b$ be a tensor field on M, and consider the expression

$$\alpha^{m\ldots c}{}_{b\ldots d}K_m{}^a + \cdots + \alpha^{a\ldots m}{}_{b\ldots c}K_m{}^c - \alpha^{a\ldots c}{}_{m\ldots d}K_b{}^m - \cdots - \alpha^{a\ldots c}{}_{b\ldots m}K_d{}^m \tag{100}$$

We regard (100) as a linear mapping from each fibre to itself. That is, if (p, $\alpha^{a\ldots c}{}_{b\ldots d}$ (α a tensor at p) is a point of B, then expression (100) defines another tensor at p, i.e., another point of B. In other words, if P is a point of B ($p = \pi(P)$), (100) defines an element of the vector space $\pi^{-1}(p)$. That is (100) defines, for each point P of B, a contravariant vector (in B) at P. Thus, we obtain a contravariant vector field on B, which we write $\overline{K}^a$. Thus, we have – a clearly linear – mapping from tensor fields $K_a{}^b$ on M to contravariant vector fields $\overline{K}^a$ on B.

Let ξ^a be a contravariant vector field on M, and let B be one of the tensor bundles over M. Set

$$\hat{\xi}^a = \tilde{\xi}^a + \overline{(\nabla\xi)}^a \tag{101}$$

Since $\vec{\pi}_P\overline{K} = 0$ for any $K_a{}^b$, and since $\tilde{\xi}^a$ is a lifting of ξ^a, it is clear that $\hat{\xi}^a$ is a lifting of ξ^a. Hence, given any tensor $\alpha^{a\ldots c}{}_{b\ldots d}$ at a point p of M, the integral curves of $\hat{\xi}^a$ through (p, $\alpha^{a\ldots c}{}_{b\ldots d}$) defines a tensor at each point of the integral curve of ξ^a trough p. This method of carrying tensors along integral curves is called *parallel transport*. It is clear from (52) that, if ξ^a and $\alpha^{a\ldots c}{}_{b\ldots d}$ are tensor fields on M, then $\alpha^{a\ldots c}{}_{b\ldots d}$ is invariant under parallel transport along the integral curves of ξ^a if and only if

$$\xi^m\nabla_m\alpha^{a\ldots c}{}_{b\ldots d}$$

To summarize, a vector field ξ^a on M can be lifted to $\tilde{\xi}^a$ on B. If, however, there is specified a derivative operator ∇_a on M. then a second lifting $\hat{\xi}^a$, is available.

The lifting $\hat{\xi}^a$ has an important property not shared by $\tilde{\xi}^a$. If ξ^a and η^a are vector fields on M, then, by (29), we have

$$\widetilde{(\xi^a + \mu\eta^a)} = \tilde{\xi}^a + \underset{\leftarrow}{(\pi\mu)}\tilde{\eta}^a - \overline{(\xi D(\mu))}^a \tag{102}$$

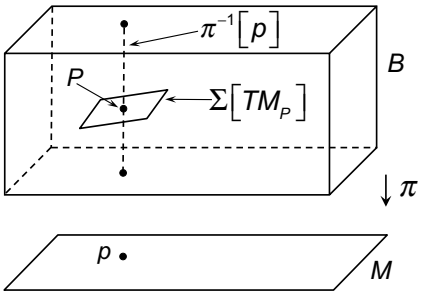
Since

$$\begin{aligned}\overline{\nabla(\xi + \mu\eta)}^a &= \overline{\nabla\xi}^a + \overline{(\mu\nabla\eta)}^a + \overline{(\eta D(\mu))} \\ &= \overline{\nabla\xi}^a + \underset{\leftarrow}{(\pi\mu)}\overline{(\nabla\eta)}^a + \overline{(\eta D(\mu))}^a\end{aligned}$$

we have immediately, from (101),

$$\widehat{(\xi^a + \mu\eta^a)} = \hat{\xi}^a + \underset{\leftarrow}{(\pi\mu)}\hat{\eta}^a \tag{103}$$

In other words, the lifting $\hat{\xi}^a$ is linear. Hence, if $p = \pi(p)$, and ξ^a and η^a are contravariant vector fields on M which coincide at p, then $\hat{\xi}^a = \hat{\eta}^a$ at P. That is $\xi^a \to \hat{\xi}^a$ defines a lifting, not only of contravariant vector fields on M, but also of contravariant vectors at a point of M.



Again, let $p = \pi(P)$. Denote by Σ the mapping from the tangent space to M at p to the tangent space to B at P which sends ξ^a to $\hat{\xi}^a$. Clearly, this mapping is linear, and satisfies $\vec{\pi}_P \cdot \Sigma =$ identity mapping on tangent space to M at p, In particular, Σ is one-to-one. The situation can now be described as follows. Let the fibres of B be $k-$dimensional. Then $\vec{\pi}_P$ sends a $k-$dimensional subspace of TB_P (tangent space to B at P) to zero, namely, the subspace of vertical vectors. On the other hand, $\Sigma[TM_P]$ is an $n-$dimensional subspace of TB_P. Evidently, every element of the $(n+k)-$dimensional vector space TB_P can be written uniquely as the sum of a vertical vector at P and an element of $\Sigma[TM_P]$. Thus, we have decomposed the vector space TB_P into two vector spaces: the vector space of vertical vectors and the vector space $\Sigma[TM_P]$. This $\Sigma[TM_P]$ is called a *horizontal subspace* of TB_P. (In general, a horizontal subspace of TB_P is a subspace such that every element of TB_P can be written uniquely as the sum of an element of the subspace and a vertical vector.)

Thus, a derivative operator on M defines, for each tensor bundle over M, a horizontal subspace of TB_P (P in B). Of course, the derivative operator is uniquely specified by the horizontal subspaces it defies. It is through these horizontal subspaces that a derivative operator on M is represented in the tensor bundles over M.

Exercise 149. Let $\gamma : I \to M$ be a curve, and let ξ^a and ξ'^a be the two vector fields on M both of which have γ as internal curve. Show that parallel transport along γ using ξ^a is identical with parallel transport along γ using ξ'^a. (Hence, parallel transport along a single curve, with no vector field, is well-defined.)

Exercise150. Consider R^2 with metric with components, in the natural chart, $g_{11} = g_{22} = 1$, $g_{12} = g_{21} = 0$. Consider parallel transport with the corresponding derivative operator. Show that the result of parallel transport of any tensor about a closed curve leaves that tensor invariant.

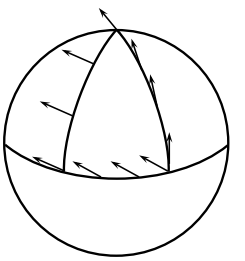
Exercise 151. In what sense does a derivative operator on M define an "inverse" of $\tilde{\vec{\pi}}$?

Exercise 152. Let ∇_a and ∇'_a be derivative operators on M, and ξ^a a contravariant vector field on M. Let $\hat{\xi}^a$ and $\hat{\xi}'^a$ denote the liftings of ξ^a defined by ∇_a and ∇'_a, respectively. Prove that $\hat{\xi}'^a = \hat{\xi}^a + \overline{K}^a$, where $K_a{}^b = \gamma^b{}_{ma}\xi^m$, and $\gamma^m{}_{ab}$ is the tensor field relating ∇'_a and ∇_a,

Exercise 153. Give an example of two contravariant fields ξ^a and η^a on M which coincide at p, but such that $\tilde{\vec{\xi}}^a \neq \tilde{\vec{\eta}}^a$ at P ($p = \pi(P)$).

Exercise 154. Prove that two derivative operators which define the same horizontal subspace are identical.

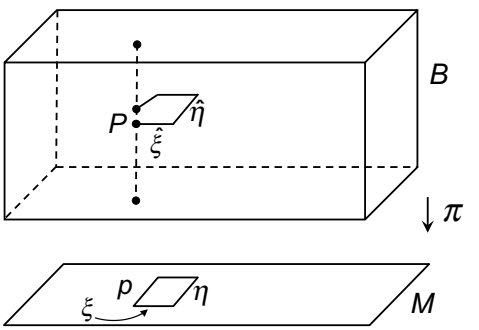
26. Riemann Tensor: Geometrical Interpretation



Let M be a manifold with derivative operator ∇_a. Then, intuitively speaking "parallel transport of a tensor about a small closed loop changes that tensor by an amount depending on the Riemann tensor". It is in this intuitive sense that the Riemann tensor represents "curvature". Parallel transport about a closed loop (small or otherwise) in the plane leaves the tensor invariant: the plane is flat. Parallel transport of a vector about a closed curve on a S^2 changes that vector: S^2 is curved. (In both cases, "parallel transport" refers to the derivative operator defined by the natural metric.)

We shell establish the following fact:

$$\mathscr{L}_{\hat{\xi}}\hat{\eta}^a - (\widehat{\mathscr{L}_{\xi}\eta^a}) = -\hat{K}^a \qquad (104)$$



where $K_a{}^b = R_{mna}{}^b\xi^m\eta^n$. Before giving the proof, we interpret (104) geometrically. The commutator in the first term on the left represents passage about an infinitesimal loop in the base space, where the "sides" of this loop are described by the infinitesimal displacements defined by ξ^a and η^a. The second term on the left in (104) represents the corresponding loop in B. The difference between these two quantities represents, therefore, the change in a tensor under parallel transport ab out the loop in M. The difference is a vertical vector, i.e., it represents a change in a tensor at p. Thus, Eqn. (104) represents the effect on a tensor under parallel transport about a small closed loop.

The proof of Eqn. (104) consists of the following calculation:

$$\begin{aligned}
&[\xi^n\nabla_n(\eta^m\nabla_m\alpha^a{}_b) - \eta^m\nabla_m(\xi^n\nabla_n\alpha^a{}_b)] \\
&- [\xi^m\nabla_m\eta^n - \eta^m\nabla_m\xi^m]\nabla_n\alpha^a{}_b = 2\xi^m\eta^n\nabla_{[m}\nabla_{n]}\alpha^a{}_b \\
&= -(\xi^m\eta^n R_{mnp}{}^a)\alpha^p{}_b + (\xi^m\eta^n R_{mnb}{}^p)\alpha^a{}_p
\end{aligned}$$

27. Geodesics

A certain class of curves, called geodesics, on a manifold with derivative operator are of particular interest.

Let M be a manifold with derivative operator ∇_a. We introduce a contravariant vector field on the tangent bundle of M, TM. If (p, ξ^a) is a point of TM, $\hat{\xi}^a$ is a contravariant vector in TM at the point (p, ξ^a). This vector field is called the *geodesic spray*. Let $\tilde{\gamma} : I \to B$ be an integral curve of the geodesic spray. Then $\gamma = \pi \cdot \tilde{\gamma}$, a curve in M, is called a *geodesic*. Thus, a geodesic is determined by a point of TM, i.e., by a contravariant vector at a point of M. The parameter t along a geodesic is sometimes called an affine parameter. An alternative definition of a geodesic is the following: a geodesic is a curve such that the tangent vector of the curve is parallel transported along the curve.

Exercise 155. Show that the geodesics on R^n with the usual metric are straight lines.

Exercise 156. Show that the geodesics on S^2 with the usual metric are great circles.

Exercise 157. Let ξ^a be a vector field on M. Show that the integral curves of ξ^a are geodesics if and only if $\xi^m \nabla_m \xi^m = 0$.

Exercise158. Let $\gamma : I \to M$ be a geodesic on M (with some derivative operator on M). Let $\mu : I' \to I$ be one function of one variable. Show that $\gamma \cdot \mu : I' \to M$ is a geodesic if and only if there are numbers a and b such that μ sends t to $at + b$.

Exercise 159. Show that, in terms of a chart, a geodesic, $x^1(t), \ldots, x^n(t)$, satisfies $d^2/dt^2 x^i(t) = -\Sigma_{j,k} \Gamma^i_{jk}(x(t)) \frac{dx^j}{dt} \frac{dx^k}{dt}$, where $\Gamma^i_{jk}(x)$ is the connection defined by the derivative operator.

Exercise160. Show that two derivative operators on M which define precisely the same geodesics are identical.

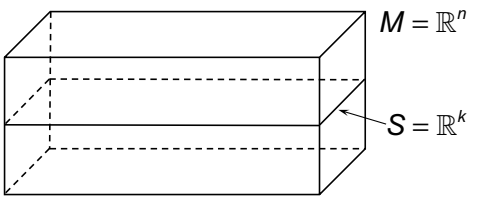
Exercise 161. Let g_{ab} be a metric on M, and let $\gamma : I \to M$ be a geodesic of the corresponding derivative operator. Show that $\xi^a \xi^b$, g_{ab} is constant along the geodesic, where ξ^a is the tangent vector of the curve.

Exercise 162. Consider a group of motions on M such that each $\Gamma_\alpha (\alpha \text{ in } G)$ takes the derivative operator to itself. Show that Γ_α takes each geodesic to another geodesic.

28. Submanifolds

Roughly speaking, a ($k-$dimensional) submanifold of a manifold M Is a ($k-$dimensional) surface in M.

Let $M = R^n$. Denote by S the collection of all points of M for which $x^{k+1} = x^{k+2} = \cdots = x^n = 0$. Clearly, this S is a $k-$dimensional hyperplane in the $n-$dimensional Euclidean space M. This is one example of a submanifold. More generally, a submanifold is a subset which, locally, looks like the example above.

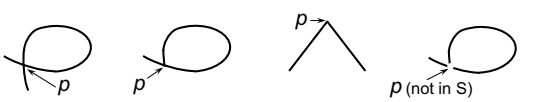


Let M be an $n-$dimensional manifold. A subset S of M is called a $k-$dimensional *submanifold* if, for each point p of S, there is a chart $(U\varphi)$ of M, with p in U, such that $\varphi[U \cap S]$ consists precisely of the points of $\varphi[U]$ with $x^{k+1} = \cdots = x^n = 0$. In other words, S is a $k-$dimensional submanifold if, for each point p of S, there is a chart containing p such that S intersects U in precisely the points of U at which the last $(n-k)$ coordinates all vanish. The number $(n-k)$ is called the *co-dimension* of the sub manifold S.

Example 65. Let S be the subset of R^n consisting of points with $(x^1)^2 + \cdots + (x^n)^2 = 1$. Then S is an $(n-1)-$dimensional submanifold of R^n. The co-dimension is one.

Example 66. Let $M = M' \times M''$. Fix a point p'' of M'', and let S consist of all points of M of the form $(p',\ p'')$, Then S is a submanifold of M, of dimension n' (= dimension of M'), and co-dimension n'' (= dimension of M'').

Example 67. The four figures below represent subset s of R^2. The first three are not submanifolds, for they do not satisfy the required condition at p. The last is a submanifold (dimension 1).



In the last figure, the "intersection point" p is not a point of S, and so no conditions need be satisfied near p. But, for each point of S, the submanifold condition is satisfied.

Example 68. The image of a cross section of a vector bundle is a sub manifold of the bundle.

Example 69. Every open subset of a manifold is an ($n-$dimensional) submanifold. Every point of a manifold is a (0−dimensional) submanifold.

Let S be a submanifold of M, so S is a set. We introduce some charts on S. If (U, φ) is a chart on M, with U S consisting of the points of U with all the last $(n-k)$ coordinates zero, then $x^1, \ldots, x^k$ are functions on $U \cap S$. Thus, $U \cap S$ is the subset of S, and $x^1, \ldots, x^k$ are the coordinates on this subset. This is a chart. (Exercise 163. Verify that these charts satisfy $M1 - M4$.) Thus, a $k-$dimensional submanifold S of M itself has the structure of a $k-$dimensional manifold. A submanifold of M represents another manifold which "sits inside" M.

Let S be a sub manifold of M, and let $\psi : S \to M$ be the identity mapping (a point of S, which is also a point of M, is taken by ψ to that point of M). Since S has the structure of a manifold, we can consider ψ as a mapping of manifolds. It is clear from the definitions that ψ is one-to-one, is smooth, and that $\vec{\psi}_p$ (p in S) is one-to-one). (ψ is one-to-one because it is the identity. ψ is smooth because the definition of the manifold structure of S gives precisely the conditions for the characterization of a smooth mapping in terms of charts. $\vec{\psi}_p$ is one-to-one because, if (U, φ) is a chart on M which defines a corresponding chart on S, and if ξ^a is a vector in S at p with components $(\tilde{\xi}^1, \ldots, \tilde{\xi}^k)$, then $\vec{\psi}_p$, ξ^a is the vector at p in M with components $(\tilde{\xi}^1, \ldots, \tilde{\xi}^k, 0, \ldots, 0)$.)

There are no suppresses here. A submanifold is a smooth surface which does not come back near it itself, and, in particular, does not intersect itself. Such a subset inherits a manifold structure from M, and the mapping from S to M becomes smooth.

Exercise 164. Find an example in which the intersection of two submanifolds is not a submanifold: in which the union of two submanifolds is not a submanifold.

Exercise 165. Let $\psi : M \to M'$ be smooth, and let S be a sub manifold of M. Find an example in which $\psi[S]$ is not a submanifold of M'.

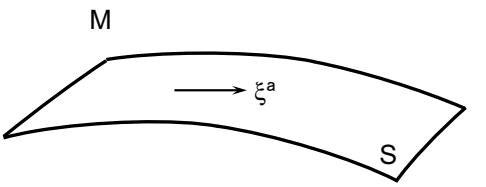
Exercise 166. Let S be a sub manifold of M, T a sub manifold of S (so T, as a subset of S, is also a subset of M). Show that T is a submanifold of M.

Exercise 167. Let $\gamma : I \to M$ be a curve with nowhere vanishing tangent vector. Show that, for each t_0 in I, there is an interval I' containing t_0 such that $\gamma[I']$ is a sub manifold of M. Find a counterexample to show this is false without the condition "with nowhere vanishing tangent vector". (Hint: Use third figure of Example 67.)

Example 70. Each fibre is a sub manifold of a vector bundle.

29. Tangents and Normals to Submanifolds

Let $\psi : S \to M$ be a $k-$dimensional submanifold of an $n-$dimensional manifold M. In this section, we relate the tensors on S to the tensors on M. [For the purposes of this section, it is convenient to regard S as a manifold separate from M, and not as a subset of M.]



Fix, once and for all, a point p' of S, and set $p = \psi(p')$, a point of M. If ξ'^a is any contravariant vector in S at p', then $\xi^a = \vec{\psi}_{p'}\xi'^a$ is a contravariant vector in M at p. A vector ξ^a such that $\xi^a = \vec{\psi}_{p'}\xi'^a$ for some ξ'^a will be said to be *tangent* to S. [Regarding S as a subset of M, this is the usual, intuitive, geometrical notion of tangency.] Evidently, the contravariant vectors at p tangent to S form a $k-$dimensional subspace of the $n-$dimensional tangent space to M at p. A contravariant vector μ_a in M at p will be said to be *normal* to S if $\xi^a\mu_a = 0$ for all ξ^a tangent to S. The normal vectors to S form an $(n-k)-$dimensional subspace of the cotangent space to M at p. Clearly, ξ^a in M at p is tangent to S if and only if $\xi^a\mu_a = 0$ for every μ_a normal to S. Thus, a $k-$dimensional submanifold S of M defines, at each point of $\psi[S]$, a $k-$dimensional subspace of the tangent space to M and $(n-k)-$dimensional subspace of the cotangent space to M.

More generally, let $\alpha^{a\ldots c}{}_{b\ldots d}$ be any tensor in M at p. A contravariant index of α, e.g., "c", will be said to be *tangent* to S if $\mu_c\alpha^{a\ldots c}{}_{b\ldots d} = 0$ for all μ_c at p normal to S. Similarly, a covariant index of α, e.g. "b", will be said to be *normal* to S if $\xi^b\alpha^{a\ldots c}{}_{b\ldots d} = 0$ for all ξ^a at p tangent to S.

Exercise 168. Verify that, for "a" in $\alpha^{a\ldots c}{}_{b\ldots d}$ tangent to S, and "a" in $\beta^{p\ldots q}{}_{m\ldots a}$ normal to S, $\alpha^{a\ldots c}{}_{b\ldots d}\beta^{p\ldots q}{}_{m\ldots a} = 0$.

Exercise 169. Prove that μ_a in M at p is normal to S if and only if $\underset{\leftarrow}{\psi}_p\,\mu_a = 0$.

Exercise 170. Let f be a scalar field on M such that f is constant on $\psi[S]$. Prove that, at each point of $\psi[S]$, $D(f)_a$ is normal to S.

Exercise 171. Let $\gamma : I \to M$ be a curve in M which remains in $\psi[S]$. Prove that, for each t, the tangent vector to γ is tangent to S.

30. Metric Submanifolds

Let $\psi : S \to M$ be a submanifold of M. [We continue to regard a submanifold S as separate from M, and not a subset.] Suppose we have a metric g_{ab} on M. Then, since we can raise and lower indices of tensors on M, additional structure is induced on M and S.

Set $h'_{ab} = \underset{\leftarrow}{\psi} g_{ab}$, a symmetric tensor field on S. If this h'_{ab} on S has an inverse, i.e., if h'^{ab} is a metric on S, then S will be said to be a *metric* sub manifold.

Example 71. Let $S = R^1$ (coordinate x), $M = R^2$, and let $\psi : S \to M$ be $\psi(x) = (x, x)$. Let g_{ab} be the metric on M whose components, in this coordinate system, are $g_{11} = 1$, $g_{22} = -1$, $g_{12} = g_{21} = 0$. Then $\underset{\leftarrow}{\psi} g_{ab} = 0$, so S is not a metric submanifold.

We now prove the following (not very important, but illustrative) result: S is a metric sub manifold if and only if, for each point p of $\psi[S]$, the only contravariant vector ξ^a at p with ξ^a tangent to S and $\xi^a = g_{ab}\xi^b$ normal to S is $\xi^a = 0$. To prove this, note that $\underset{\leftarrow}{\psi}_{p'} g_{ab}$ is not invertible if and only if it has a nullspace, i.e., if and only if $\eta'^a \underset{\leftarrow}{\psi}_{p'} g_{ab} = 0$ for some nonzero vector η'^a in S at p' (where $p = \psi(p')$). But, if η'^a satisfies these conditions, then $\xi^a = \vec{\psi}_{p'} \eta'^a$ is tangent to S, while $\underset{\leftarrow}{\psi}_{p'} g_{ab}\xi^b = \underset{\leftarrow}{\psi}_{p'} (g_{ab} \vec{\psi}_{p'} \eta'^a) = \eta'^b \underset{\leftarrow}{\psi}_{p'} g_{ab} = 0$, so ξ_a is normal to S. Conversely, if ξ^a is tangent to S, with ξ_a normal to S, then the η'^a in S at p' such that $\vec{\psi}_{p'} \eta'^a = \xi^a$ is in the nullspace of $\underset{\leftarrow}{\psi}_{p'} g_{ab}$, so S is not a metric submanifold.

Suppose for a moment that g_{ab} were positive-definite. Then, if S were not a metric submanifold, there would exist a nonzero ξ^a tangent to S with Σ_a normal to S. But this would imply $\xi^a\xi_a = 0$, which is impossible for a nonzero vector and positive-definite metric. We conclude that every submanifold of a manifold with positive-definite metric is a metric submanifold.

Let g_{ab} be a metric on M, $\psi : S \to M$ a metric submanifold of M. Let $\alpha^{a\ldots c}{}_{b\ldots d}$ be a tensor in M at a point p of $\psi[S]$. We say that an index of α is *normal* to S if, when that index is lowered (if necessary; with the metric) it is normal to S (as in Sect. 29). Similarly, an index of α is *tangent* to S if, when that index is raised, it is tangent to S. An index of a tensor cannot be simultaneously normal and tangent unless the tensor itself vanishes.

Let S be a $k-$dimensional metric submanifold of M. What do we get at a point p of $\psi[S]$? We have the $k-$dimensional subspace of the tangent space at p consisting of vectors tangent to S. We also have the $(n-k)-$dimensional subspace of the cotangent space of p consisting of vectors normal to S. But, since M has a metric, we can raise the indices of the vectors in the latter subspace. Thus, we obtain an $(n-k)-$dimensional subspace of the tangent space at p consisting of vectors normal to S. These two subspace of the tangent space at p intersect only at the origin. Furthermore, their dimensions – k and $(n-k)$ – add up to n, the dimension of the tangent space at p. Thus, we have decomposed the tangent space at p into the direct sum of two subspaces. It should be possible to introduce projections into the two subspaces, etc. We now do this.

As usual, fix p' in S, and set $p = \psi(p')$. Since $\underset{\leftarrow}{\psi}\, g_{ab}$ is a metric on S, the inverse metric, h'^{ab}, exists (on S). Set $h^{ab} = \vec{\psi}_{p'} h'^{ab}$, a symmetric tensor in M at p. It is immediate from the definition that both indices of h^{ab} are tangent to M. We raise and lower indices of h^{ab} using the metric g_{ab} of M. We now prove the following: a contravariant vector ξ^b at p is tangent to S if and only if $\xi^a = h^a{}_b \xi^b$. The "if" part is clear. To prove the converse,

let ξ^a be tangent to S, so $\xi^a = \vec{\psi}_{p'}\, \xi'^a$. then

$$\begin{aligned}\underset{\leftarrow}{\psi}_{p'}\xi_a &= \underset{\leftarrow}{\psi}_{p'}\,(g_{ab}\xi^b) = \underset{\leftarrow_{p'}}{\psi}\,(g_{ab}\vec{\psi}\xi'^b) = \xi'^b \underset{\leftarrow}{\psi}\, g_{ab}\\ &= h'_{ab}\xi'^b\end{aligned}$$

Hence,

$$\begin{aligned}\xi^a &= \vec{\psi}_{p'}\xi'^a = \vec{\psi}_{p'}\,(h'^{ab}\, h'_{bc}\xi'^c) = \vec{\psi}_{p'}\,(h'^{ab}\underset{\leftarrow}{\psi}_{p'}\xi_b)\\ &= (\vec{\psi}_{p'}\, h'^{ab})\,\xi_b = h^a{}_b\xi^b\end{aligned}$$

Similarly, a covariant μ_a in M at p is normal to S if and only if $\mu_a\, h^a{}_b = 0$. [Proof: The "only if" is clear. Let $\mu_a\, h^a{}_b = 0$ Then, for every ξ^a tangent to S, $\mu_a\,\xi^a = \mu_a\,(h^a{}_b\,\xi^b) = 0$. So, μ_a is normal to S.] By the same argument, we see that an index, e.g., "a". of a tensor $\alpha^{\cdots a\cdots}$ at p is tangent to S if and only if

$$h^a{}_b\alpha^{\cdots b\cdots} = \alpha^{\cdots a\cdots} \tag{105}$$

and normal to S if and only if

$$h^a{}_b\alpha^{\cdots b\cdots} = 0 \tag{106}$$

In particular, $h^a{}_b h^b{}_c = h^a{}_c$. Hence, setting $k_{ab} = g_{ab} - h_{ab}$, we have

$$h^a{}_b k^b{}_c = h^a{}_b(\delta^b{}_c - h^b{}_c) = h^a{}_c - h^a{}_c = 0 \tag{107}$$

and

$$\begin{aligned}k^a{}_b k^b{}_c &= (\delta^a{}_b - h^a{}_b)(\delta^b{}_c - h^b{}_c) = \delta^a{}_c - 2h^a{}_c + h^a{}_c\\ &= \delta^a{}_c - h^a{}_c = k^a{}_c\end{aligned} \tag{108}$$

Evidently, ξ^a satisfies $\xi^b\, k^a{}_b = 0$ if and only if ξ^a satisfies $\xi^a = \xi^b\, h^a{}_b$. Thus, reversing "tangent" and "normal", statements about h_{ab} become statement about k_{ab}. We summarize:

Theorem 21. Let g_{ab} be a metric on M, $\psi : S \to M$ a metric submanifold of M, and $p = \psi(p')$ a point of M. Then there are tensors h_{ab} and k_{ab} in M at p with the properties:

1. $h_{ab} + k_{ab} = g_{ab}$
2. Both indices of h_{ab} are tangent to S; both indices of k_{ab} are normal to S.
3. An index "a" of a tensor $\alpha^{\cdots a\cdots}$ at p is tangent to S if and only if $h^a{}_b\alpha^{\cdots b\cdots} = \alpha^{\cdots a\cdots}$, which, in turns, holds if and only if $k^a{}_b\alpha^{\cdots b\cdots} = 0$.
4. An index "a" of a tensor $\alpha^{\cdots a\cdots}$ at p is normal to S if and only if $k^a{}_b\alpha^{\cdots b\cdots} = \alpha^{\cdots a\cdots}$, which in turn, holds if and only if $h^a{}_b\alpha^{\cdots b\cdots} = 0$.
5. $h_a{}^b h_b{}^c = h_a{}^c$, $k_a{}^b k_b{}^c = k_a{}^c$ (projections operators).

A given index of a tensor at p may, of course, be neither tangent nor normal to S. But Theorem 21 has the useful consequence that any tensor may be decomposed into tensors all of whose indices are either tangent or normal. We might as well take all indices contravariant. We do the case with two indices as an example: the $n-$index case should be clear:

$$\begin{aligned}\alpha^{ab} &= \alpha^{mn}\delta^a{}_m\delta^b{}_n = \alpha^{mn}(h^a{}_m + k^a{}_m)(h^b{}_n + k^b{}_n)\\ &= \alpha^{mn}h^a{}_m h^b{}_n + \alpha^{mn}k^a{}_m h^b{}_n = \alpha^{mn}h^a{}_m k^b{}_n\\ &+ \alpha^{mn}k^a{}_m k^b{}_n\end{aligned} \tag{109}$$

To summarize, the tensors at a point of M on a metric submanifold have the nicest decomposition one could have possible asked for.

Exercise 172. what goes wrong with above if S is not a metric submanifold?

Exercise . 173 Prove that $h^a{}_a = k$ (=dimension of S); that $k^a{}_a = (n-k)$
Exercise 174. Consider Example 65. Let $M = R^n$ have metric with components $g_{11} = \ldots = g_{nn} = 1$, all others zero. Show that S is a metric sub manifold, and find h_{ab} and k_{ab}.
Exercise 175. Show that, in co-dimension one, $k_{ab} = \pm\xi_a\,\xi_b$ where ξ_a is a normal vector to S. When will the plus sign be applicable?
Exercise 176. Let the submanifold S have dimension one, so S is the image of a curve. Show that $h_{ab} = \alpha\eta_a\eta_b$, where α is a scalar and η_a is the tangent to the curve.
Exercise 177. Find an example of a metric submanifold S such that the metric h'_{ab} on S is indefinite.
Exercise 178. Find h_{ab} and k_{ab} for a manifold of co-dimension zero.
Exercise 179. Define the signature of h_{ab} and k_{ab}, and prove a theorem relating these signatures to that of g_{ab}.

31. Tensor Fields and Derivatives on Submanifolds

Let M be a manifold, and S a subset of M such that S is a $k-$dimensional sub manifold. (Note that we are now taking a submanifold as a subset. We shall allow ourselves to switch from one interpretation to the other (equivalent) one.) A *tensor field on* the submanifold S consists of an assignment, to each point p of S, a tensor (in M) at p, where all the tensor at various points have the same index structure. Thus, if B is the appropriate tensor bundle over M, a tensor field on S is a mapping $\Lambda : \ S \rightarrow B$ such that $\pi \cdot \Lambda =$ identity on S. (Example: When $S = M$, tensor fields are cross sections.) A tensor field on S is said to be *smooth* if the corresponding mapping of manifolds, Λ, is smooth.

Smoothness can be expressed in terms of components. Let (U, φ) be a chart on M such that at $U \cap S$ consisting precisely of the points of U with coordinates $x^{k+1} = \ldots = x^n = 0$. Let $\alpha^{a\ldots c}{}_{b\ldots d}$ be a tensor field on the submanifold S. Then the components of $\alpha^{a\ldots c}{}_{b\ldots d}$ (with respect to this chart) are functions of the k variables $x^1, \ldots, x^k$ (for $\alpha^{a\ldots c}{}_{b\ldots d}$ is only defined on S, were the last $(n-k)$ coordinates vanish). Evidently, $\alpha^{a\ldots c}{}_{b\ldots d}$ is smooth if and only if these functions of k variables are C^∞ for every such chart. (Exercise 180. Prove this.)

Clearly, addition, outer products, contraction, and index substitution on smooth tensor fields on the submanifold S result in smooth tensor fields on the submanifold S. As usual, we normally omit the adjective "smooth". (Differential geometry is the study of smoothness.)

There is a chance of confusion in the terminology here. If we consider S as a manifold in its own right, then we have tensor fields on this manifold. These are different from tensor fields on the submanifold S, as defined above. We keep this distinction by calling the former "tensor fields on the manifold S," and the latter "tensor fields on the submanifold S." (Exercise 181. Find a natural, one-to-one correspondence between contravariant tensor fields on the manifold S and contravariant tensor fields on the submanifold S all of whose indices are tangent to S.)

The remarks above apply, of course, to submanifolds in general, i.e., without further structure. The next logical step would be to introduce the notion of a derivative operator on the submanifold S, and then prove that derivative operator on M induces a natural derivative operator on the submanifold S. This is straightforward, devoid of surprises, and ... not particularly useful. We shall, therefore, omit the discussion of these intermediate structures, and proceed directly to the most common and important case: metric submanifolds. (Exercise 182. Define a derivative operator on the submanifolds S. Show that a derivative operator on M indices a natural derivative operator on each submanifold S.)

Let M be a manifold, g_{ab} a metric on M, and S, a subset of M, a $k-$dimensional metric sub manifold of M. Then g_{ab}, h_{ab}, and k_{ab} (Sect. 30) are tensor fields on the submanifold S. (Exercise 183. Prove smooth.) Let ∇_a be the derivative operator (on M) associated with g_{ab}. Finally, let $\alpha^{b\ldots c}{}_{r\ldots s}$ be any (smooth) tensor field on the submanifold S. It is clear from the characterization in terms of charts on the previous page that there exists a tensor field $\tilde{\alpha}^{b\ldots c}{}_{r\ldots s}$ on M such that, at points of S, $\tilde{\alpha}^{b\ldots c}{}_{r\ldots s}$ coincides with $\alpha^{b\ldots c}{}_{r\ldots s}$. (In fact, it is only immediately clear that the is true locally. That's all we need. It is true as stated, i.e., globally.) We shall call $\tilde{\alpha}^{b\ldots c}{}_{r\ldots s}$ an extension of $\alpha^{b\ldots c}{}_{r\ldots s}$ to M. Evidently, if $\tilde{\alpha}^{b\ldots c}{}_{r\ldots s}$ and $\tilde{\alpha'}^{b\ldots c}{}_{r\ldots s}$ are two extensions of $\alpha^{b\ldots c}{}_{r\ldots s}$ then $\tilde{\alpha}^{c\ldots c}{}_{r\ldots s} - \tilde{\alpha'}^{b\ldots c}{}_{r\ldots s}$ vanishes on S, so

$$\tilde{\alpha}^{b\ldots c}{}_{r\ldots s} - \tilde{\alpha'}^{b\ldots c}{}_{r\ldots s} = \mu\beta^{b\ldots c}{}_{r\ldots s} + \ldots + \nu\gamma^{b\ldots c}{}_{r\ldots s} \tag{110}$$

where all fields are on M, and the scalar fields vanish on S. (Exercise 184. By using

components, establish the existence of such an expansion (110).) Now, let p be a point of S, ξ^a a vector at p tangent to S. We define $\xi^a D_a \alpha^{b\ldots c}{}_{r\ldots s}$ by the equation

$$\xi^a \, D_a \, \alpha^{b\ldots c}{}_{r\ldots s} = \xi^a \, \nabla_a \tilde{\alpha}^{b\ldots c}{}_{r\ldots s} \tag{111}$$

where $\tilde{\alpha}^{b\ldots c}{}_{r\ldots s}$ is an extension of $\alpha^{b\ldots c}{}_{r\ldots s}$. The right side of (111) is independent of the choice of extension, for, using (110),

$$\xi^a \nabla_a \tilde{\alpha}^{b\ldots c}{}_{r\ldots s} = \xi^a \nabla_a \tilde{\alpha}'^{b\ldots c}{}_{r\ldots s} + \mu \xi^a \nabla_a \beta^{b\ldots c}{}_{r\ldots s} + \beta^{b\ldots c}{}_{r\ldots s} \xi^a \nabla_a \mu$$
$$+ \cdots + \nu \xi^a \nabla_a \gamma^{b\ldots c}{}_{r\ldots s} + \gamma^{b\ldots c}{}_{r\ldots s} \xi^a \nabla_a \nu = \xi^a \nabla_a \tilde{\alpha}'^{b\ldots c}{}_{r\ldots s}$$

Thus, $xi^a \, D_a \alpha^{b\ldots c}{}_{r\ldots s}$ has a "natural" definition only for ξ^a tangent to S. (Intuitively, for tensor fields only defined on S, the derivative can only be taken, meaningfully, along directions along S.) For ξ^a normal to S, set $\xi^a \, D_a \, \alpha^{b\ldots c}{}_{r\ldots s} = 0$. Thus (since every vector is a unique linear combination of a vector tangent to S and a vector normal to S) $\xi^a \, D_a \, \alpha^{b\ldots c}{}_{r\ldots s}$ is defined for all ξ^a, and is linear in ξ^a. Thus, we obtain a tensor field $D_a \, \alpha^{b\ldots c}{}_{r\ldots s}$ on the sub manifold S. The index "a" in $D_a \, \alpha^{b\ldots c}{}_{r\ldots s}$ is, clearly, tangent to S. (This says the same thing: only derivatives in directions tangent to S are meaningful. By fiat, we set derivatives in normal directions equal to zero.) The various properties of ∇_a carry over immediately (e.g., from (111)) to D_a. Thus, we have the Leibnitz rule,

$$D_a \, (\alpha^{\cdots}{}_{\ldots} \beta^{\cdots}{}_{\ldots}) = \alpha^{\cdots}{}_{\ldots} \, D_a \, \beta^{\cdots}{}_{\ldots} + \beta^{\cdots}{}_{\ldots} \, D_a \, \alpha^{\cdots}{}_{\ldots} \tag{112}$$

additivity,

$$D_a \, (\mu^{\cdots}{}_{\ldots} + \nu^{\cdots}{}_{\ldots}) = D_a \, \mu^{\cdots}{}_{\ldots} + D_a \, \nu^{\cdots}{}_{\ldots} \tag{113}$$

etc.

To summarize, if S is a metric submanifold of M, then with any tensor field, $\alpha^{b\ldots c}{}_{r\ldots s}$, on the sub manifold S, there is associated another tensor field, $D_a \, \alpha^{b\ldots c}{}_{r\ldots s}$, on the submanifold S, where the index "a" is tangent to S. This D_a is additive and satisfies the Leibnitz rule, and the application of D_a commutes with contraction and index substitution. [Note: $D_a = h_a{}^m \nabla_a$].

We establish one further important property of this D_a. Let $\psi : \; S \to M$ be a metric submanifold of M. Then, if $\alpha_{b\ldots c}$ is a tensor field on the submanifold S all of whose indices are tangent to S, then $\alpha'_{b\ldots c} = \underset{\leftarrow}{\psi} \, \alpha_{b\ldots c}$ is a tensor field on the manifold S. This is, evidently, a one-to-one correspondence between tensor fields on the manifold S and the tensor fields on the submanifold S all of whose indices are tangent to S. Now,

$$D'_a \alpha'_{b\ldots c} = \underset{\leftarrow}{\psi} \, (D_a \, \alpha_{b\ldots c})$$

defines an operator D'_a on covariant tensor fields on the manifold S. It is clear from the summary above that this D'_a is, in fact, a derivative operator on the manifold S. We prove that this D'_a is precisely the derivative operator on the manifold S defined by the metric h'_{ab} on the manifold S. The proof is easy:

$$D'_a \, h'_{bc} = D'_a (\underset{\leftarrow}{\psi} \, g_{bc}) = \underset{\leftarrow}{\psi} \, (D_a \, g_{bc}) = \underset{\leftarrow}{\psi} (h_a{}^m \nabla_m g_{bc}) = \underset{\leftarrow}{\psi} (0) = 0$$

Thus, instead of talking of tensor fields on the manifold S, we can talk of tensor fields on the submanifold S all of whose indices are tangent to S. If $\alpha_{b\ldots c}$ is such a tensor field on the submanifold S, then

$$h_b{}^{m\ldots} h_c{}^n \, D_a \, \alpha_{m\ldots n} \tag{114}$$

is another. Thus, we have a derivative operator on the manifold S, and this turns out to be precisely the derivative operator defined by the metric on the manifold S. Quite generally, we can ignore the manifold S: tensors and tensor operations on the manifold S can be

replaced by certain tensors (those whose indices are tangent to S) and tensor operations on the sub manifold S.

Example 72. Let the curve $\gamma : I \to M$ be a metric submanifold of M. Then a tensor field $\alpha^{b\ldots c}{}_{r\ldots s}$ on the submanifold $\gamma[I]$ satisfies $D_a \alpha^{b\ldots c}{}_{r\ldots s} = 0$ if and only if $\alpha^{b\ldots c}{}_{r\ldots s}$ is parallel transported along the curve γ.

Exercise 185. Considering $\delta^b{}_c$ as a tensor field on the submanifold S, show that $D_a \delta^b{}_c = 0$.

Exercise 186. Prove that $D_a g_{bc} = 0$.

Exercise 187. Prove that $h_b{}^m h_c{}^n D_a h_{mn} = 0$.

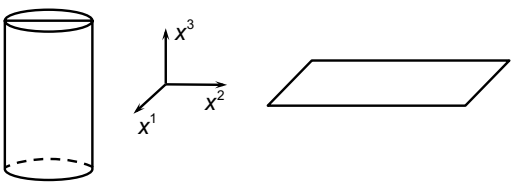
Exercise 188. Prove that D_a commutes with raising and lowering indices of tensor fields on the sub manifold S (using g_{ab}).

Exercise 189. Let ξ^a and η^a be vector fields on the submanifold S, with indices tangent to S. Define $\mathscr{L}_\xi \eta = \xi^m D_m \eta^a - \eta^m D_m \xi^a$. (Why is a definition needed?) Show that this expression satisfies all the properties of Lie derivatives of vectors.

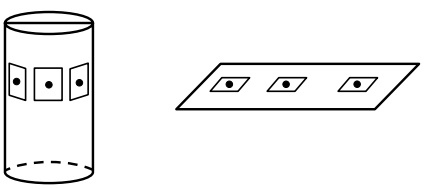
Example 73. If $S = M$, all the constructions above reduce to standard differential geometry on M.

Example 74. Let S be a submanifold of M. Then the tensors (in M) at points of S form vector bundles over S (as base space). Note that the dimensions of the fibres depend only on the dimension of M (and, of course, on the rank of the tensors under consideration), are not on the dimension of S. Now suppose S is a metric submanifold of M. Then, D_a induces a horizontal subspace of these bundles. Furthermore, each of these bundles over S is a submanifold of the corresponding bundle over M. The horizontal subspace of the two bundles are related in the obvious way.

32. Extrinsic Curvature



Let $M = R^3$ with the usual positive-definite metric g_{ab}. Consider the two submanifolds of M consisting, on the one hand, of the cylinder $((x^1)^2 + (x^2)^2 = 1)$, and, on the other hand, of the plane $(x^3 = 0)$. Since the metric is positive-definite, these are metric submanifolds. The induced metric h'_{ab} on each submanifold is flat (has vanishing Riemann tensor). (Intuitively, a piece of paper in R^3 can form either a plane or a cylinder.) But, clearly, the cylinder is "curved" in a way that the plane is not. This "curvature" is not an intrinsic property of the geometry of the submanifold, but, rather, expresses the "way in which the submanifold is embedded in M". The embedding of a metric submanifold S in M is described by a certain tensor field on the submanifold S called the extrinsic curvature. It is the extrinsic curvature, and not the intrinsic (local) geometry which distinguishes the cylinder in R^3 from the plane in R^3.



Let S be a metric sub manifold of M. The discussion of Sect. 31 suggests that the quantities $D_a h_{bc}$ and $D_a k_{bc}$ (of course, $D_a h_{bc} = D_a k_{bc} = 0$, so only one needs be considered) should be investigated. Intuitively, $D_a h_{bc}$ represents the "rate of change of the tangent plane to S along S". We might expect this quantity to be zero for the plane (where "all the tangent planes to S at points of S are parallel"), and nonzero for the cylinder (where "the tangent plane to S changes orientation as we move about S").

Thus, $D_a h_{bc}$ does, intuitively, represents the "bending of S in M" – the quantity which distinguishes the plane from the cylinder. We proceed to the study of $D_a h_{bc}$.

The first step in studying any tensor field on the sub manifold S is to project the indices tangent and normal to S (See (109).). this simplification is always possible – and always fruitful. First note that

$$\begin{aligned} h_b{}^m h_c{}^n D_a h_{mn} &= h_b{}^m D_a (h_c{}^n h_{mn}) - h_b{}^m h_{mn} D_a h_c{}^n \\ &= h_b{}^m D_a h_{cm} - h_b{}^n D_a h_{cn} = 0 \end{aligned}$$

Furthermore,

$$\begin{aligned} k_b{}^m k_c{}^n D_a h_{mn} &= k_b{}^m D_a (k_c{}^n h_{mn}) - k_b{}^m h_{mn} D_a k_c{}^n \\ &= k_b{}^m D_a(0) - 0\, D_a k_c{}^n = 0 \end{aligned}$$

hence, only the "mixed piece" remains:

$$h_b{}^m k_c{}^n D_a h_{mn} = \pi_{abc} \tag{115}$$

This tensor field π_{abc} on the (metric) submanifold S is called the *extrinsic curvature* of S. Note that the first two indices of π_{abc} are tangent to S, the last index normal to S. (One can convince himself intuitively that this is the correct index structure to describe the "bending" of S in M.)

From the results above,

$$\begin{aligned} D_a h_{bc} &= \delta_b{}^m \delta_c{}^n D_a h_{mn} = (h_b{}^m + k_b{}^m)(h_c{}^n + k_c{}^n) D_a h_{mn} \\ &= (k_b{}^m h_c{}^n + h_b{}^m k_c{}^n)\, D_a h_{mn} = 2\pi_{a(bc)} \end{aligned}$$

Hence, $D_a k_{bc} = -2\pi_a(bc)$. So, the extrinsic curvature completely determines $D_a\, k_{bc}$ and $D_a k_{bc}$

There is one additional symmetry of π_{abc}, which we now establish. Let ξ^a be a vector field on the submanifold S, normal to S. Then

$$\begin{aligned}\pi_{[ab]c}\xi^c &= \xi^c k_c{}^n h_{[b}{}^m D_{a]} h_{mn} = \xi^n h_{[b}{}^m D_{a]} h_{mn} = h_{[b}{}^m D_{a]}(\xi^n h_{mn}) \\ &- h_{mn} h_{[b}{}^m D_{a]}\xi^n = -h_{[b}{}^n D_{a]}\xi^n = -h_a{}^m h_b{}^n \nabla_{[m}\tilde{\xi}_{n]}\end{aligned} \tag{116}$$

Let $\tilde{\xi}_n$ be an extension of ξ^n. We now "pull back" the right side of (116) using $\underset{\leftarrow}{\psi}$ (where $\psi : S \to M$):

$$\begin{aligned}\underset{\leftarrow}{\psi}(h_a{}^m h_b{}^n \nabla_{[n}\tilde{\xi}_{n]}) &= \underset{\leftarrow}{\psi}\nabla_{[m}\tilde{\xi}_{n]} = \underset{\leftarrow}{\xi} D(\tilde{\xi})_{mn} \\ &= D(\underset{\leftarrow}{\psi}\tilde{\xi})_{mn} = 0\end{aligned}$$

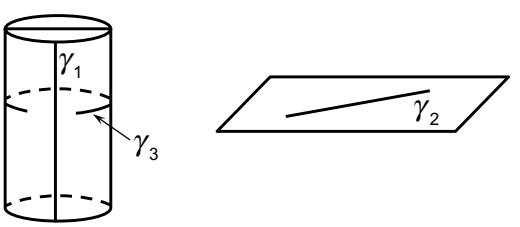
where, for the last equality, we have used the fact that ξ_a is normal to S. Thus, $\pi_{[ab]c}\xi^c$ has all indices tangent to S, and $\underset{\leftarrow}{\psi}(\pi_{[ab]c}\xi^c) = 0$: so $\pi_{[ab]c}\xi^c = 0$. But ξ^c is arbitrary (normal to S, but it makes no difference, since "c" of π_{abc} is normal to S), whence $\pi_{[ab]c} = 0$ that is to say, π_{abc} is symmetric in its first two indices. (Note, incidentally, that a nonzero tensor will never be either symmetric or antisymmetric under interchange of a pair of indices one of which is normal, the other tangent, to S.) Thus, π_{abc} has $(n-k)k(k+1)/2$ independent components.

There is a pretty little formula which gives a good geometrical picture of what the extrinsic curvature is. Consider a geodesic within the sub manifold S, so its tangent vector (which is tangent also to S) satisfies

$$h^a{}_c \eta^b D_b \eta^c = 0 \tag{117}$$

We ask for the curvature of this curve, considered as a curve in M. We have

$$\begin{aligned}\eta^b \nabla_b \eta^c &= \eta^b D_b \eta^c = k^a{}_c \eta^b D_b \eta^c = \eta^b D_b (k^a{}_c \eta^c) \\ &- \eta^b \eta^c D_b k^a{}_c = 0 - \eta^b \eta^c D_b k^a{}_c \\ &= 2\pi_{bc}{}^a \eta^b \eta^c\end{aligned} \tag{118}$$



Thus, although our curve in S is a geodesic in S, it in general has curvature in M. This curvature vector is normal to M, and given by the extrinsic curvature. Evidently, these "curvatures of geodesics" determine π_{abc} completely. Consider again the plane and cylinder. The curves γ_1 and γ_2 are geodesics in their submanifolds, and also geodesics (straight lines) in R^3. Thus, the extrinsic curvature of the plane is zero, while $\pi_{ab}{}^c \eta^a \eta^b$ for η^a tangent to γ_1 in the cylinder is zero. On the other hand, γ_3 is a geodesic in the cylinder, but not a geodesic in R^3. Hence, $\pi_{ab}{}^c \eta^a \eta^b$ for η^a tangent to γ_3 is not zero.

Thus, the extrinsic curvature represents the amount by which "curves which insist on remaining in S, but otherwise try to be as geodesic as possible, are forced to curve in order to remain in S". Why is the right side of (118) normal to S? Physical analogy: consider a point particle, free to move in R^3 except that it is constrained to lie in some surface S. Then the surface S will find it necessary to exert a force on the particle to keep it in S, and this force will be normal to S.

Exercise 190. Find the extrinsic curvature for the plane and cylinder examples, and verify explicitly Eqn. (118).

Example 75. A curve, considered as a submanifold, is a geodesic if and only if its extrinsic curvature vanishes.

Example 76. For co-dimension zero, the extrinsic curvature always vanishes.

Exercise 191. Let S be a manifold with metric h'_{ab}. Consider R^m with its natural positive-definite metric. Set $M = S \times R^m$, and let g_{ab} be the metric on M such that the induced metric on each submanifold S and each submanifold R^m is the given one. (This construction is called the *product* of manifolds-with-metric.) Consider the submanifold (a copy of S) consisting of pairs (p, x) with p in S and $x = 0$ (in R^m). Prove that its extrinsic curvature is zero. (Conclusion: The intrinsic geometry of a submanifold allows no prediction about what its extrinsic curvature will be.)

Exercise 192. Find an example of a metric submanifold whose intrinsic metric has vanishing Riemann tensor, but whose extrinsic curvature is nonzero.

Exercise 193. Find all 2-dimensional submanifolds of R^3 (with usual metric) having vanishing extrinsic curvature.

Exercise 194. Find the extrinsic curvature of S^{n-1} in R^n (usual embedding, usual metric.).

Exercise 195. Find a manifold-with-metric such that no submanifold of co-dimensional one has vanishing extrinsic curvature.

Exercise 196. Let S be a metric submanifold of M, T a metric submanifold of S. Let S have vanishing extrinsic curvature in M, and T vanishing extrinsic curvature in S. Prove that T vanishing extrinsic curvature in M.

Exercise 197. Let M be a manifold-with-metric, S a (not necessarily metric) submanifold. How much of this extrinsic curvature business goes through?

33. The Gauss-Codazzi Equations

Let S be a metric submanifold of M (with g_{ab}). There exists a set of two equations relating the geometry of M, the geometry of S, and the extrinsic curvature of S. These are called the Gauss-Codazzi equations. They turn out to be useful in numerous applications. It is perhaps not surprising that such equations should exist. Consider a two-dimensional submanifold of R^3 (with the usual metric). Intuitively, if the surface is very "wiggly" in R^3, then it would tend to have a large intrinsic Riemann tensor, while a relatively "flat" S would have a small Riemann tensor.

The first Gauss-Codazzi equation arises from the following fact from Sect. 31: there is a natural, one-to-one, correspondence between tensor fields on the manifold S and tensor fields on the submanifold S all of whose indices are tangent to S. Furthermore, the mapping from tensor fields on the submanifold S all of whose indices are tangent to S to tensor fields on the submanifold S all of whose indices are tangent to S defined by

$$h^b{}_m \cdots h^c{}_n h_d{}^p \cdots h_e{}^q D_a \alpha^{m\ldots n}{}_{p\ldots r} \tag{119}$$

corresponds precisely to the mapping from tensor fields on the manifold S to tensor fields on the manifold S defined by the derivative operator on S (associated with the metric h'_{ab} of S). Thus, the Riemann tensor of the manifold S can be obtained by a calculation in M. Specifically, if k_a is a vector field on the sub manifold S which is tangent to S, then

$$\frac{1}{2}\mathcal{R}_{abc}{}^d k_d = h_{[a}{}^m h_{b]}{}^n h_c{}^p \nabla_m (h_n{}^r h_p{}^s \nabla_r k_s) \tag{120}$$

where $\mathcal{R}_{abc}{}^d$ is the Riemann tensor of S. The idea is to expand the right side of (120):

$$\begin{aligned} \frac{1}{2}\mathcal{R}_{abc}{}^d k_d &= h_{[a}{}^m h_{b]}{}^n h_c{}^p [(\nabla_m h_n{}^r) h_p{}^s \nabla_r k_s) \\ &+ h_n{}^r (\nabla_m h_p{}^s) \nabla_r k_s + h_n{}^r h_p{}^s \nabla_m \nabla_r k_s] \end{aligned} \tag{121}$$

The first term on the right in (121) vanishes, for $h_{[a}{}^m h_{b]}{}^n \nabla_n h_n{}^r = \pi_{[ab]}{}^r = 0$. The last two terms in (121) give

$$\begin{aligned} \frac{1}{2}\mathcal{R}_{abc}{}^d k_d &= h_{[a}{}^m h_{b]}{}^r h_c{}^p (\nabla_m h_p{}^s) \nabla_r k_s + h_{[a}{}^m h_{b]}{}^r h_c{}^s \nabla_m \nabla_r k_b \\ &= h_{[b}{}^r \pi_{a]c}{}^s \nabla_r k_s + h_a{}^m h_b{}^r h_c{}^s \nabla_{[m} \nabla_{r]} k_s \\ &= h_{[b}{}^r \pi_{a]c}{}^s \nabla_r k_s + \frac{1}{2} h_a{}^m h_b{}^r h_c{}^s R_{mrs}{}^d k_d \end{aligned} \tag{122}$$

were, for the first step, we have used $h_a{}^m h_m{}^b = h_a{}^b$ (repeatedly), for the second step, we have used the definition of $\pi_{ab}{}^c$, for the third step, we have used the definition of the Riemann tensor, $R_{abc}{}^d$, of the manifold M. Finally, we modify the first term on the far right side of (122) as follows:

$$\begin{aligned} h_{[b}{}^r \pi_{a]c}{}^s \nabla_r k_s &= h_{[b}{}^r \pi_{a]c}{}^s \nabla_r (h_s{}^m k_s) = h_{[b}{}^r \pi_{a]c}{}^s k_m \nabla_r h_s{}^m \\ &+ h_{[b}{}^r \pi_{a]c}{}^s h_s{}^m \nabla_r k_m = h_{[b}{}^r \pi_{a]c}{}^s k_m \nabla_r h_s{}^m = \pi_{c[a}{}^s \pi_{b]ms} k^m \end{aligned}$$

where, in the third step, we have used the fact that $\pi_{ac}{}^s h_s{}^m = 0$ (a normal index contracted with a tangent index). Thus, (122) becomes

$$\frac{1}{2}\mathcal{R}_{abc}{}^d k_d = \pi_{c[a}{}^s \pi_{b]ms} k^m + \frac{1}{2} h_a{}^m h_b{}^r h_c{}^s R_{mrs}{}^d k_d$$

Using the fact that k_c is arbitrary (tangent to S), we obtain the first *Gauss-Codazzi equation*:

$$\mathcal{R}_{abcd} = 2\,\pi_{c[a}{}^{m}\pi_{b]dm} + h_a{}^m h_b{}^n h_c{}^p h_d{}^q R_{mnpq} \tag{123}$$

This equation states that the Riemann tensor of the sub manifold S (considered as a manifold-with-metric in its own right) is equal to the Riemann tensor of M, with its indices all projected tangent to S, plus a term involving the extrinsic curvature. Intuitively, the "intrinsic wiggliness of M" (described by the Riemann tensor of M) and the "wiggliness of the embedding of S in M" (described by the extrinsic curvature) combine to give the "intrinsic wiggliness of S" (described by the Riemann tensor of S).

To obtain the second Gauss-Codazzi equation, we proceed as follows:

$$\begin{aligned} &h_{[a}{}^m h_{b]}{}^n h_c{}^p k_q{}^d \nabla_m \pi_{np}{}^q = h_{[a}{}^m h_{b]}{}^n h_c{}^p h_q{}^d \nabla_m [h_n{}^r h_p{}^s k^{qt} \nabla_r h_{st}] \\ &= h_{[a}{}^m h_{b]}{}^n h_c{}^p h_q{}^d [(\nabla_m h_n{}^r h_p{}^s k^{qt} \nabla_r h_{st} \\ &+ h_n{}^r (\nabla_m h_p{}^s) k^{qt} \nabla_r h_{st} + h_n{}^r h_p{}^s (\nabla_m k^{qt}) \nabla_r h_{st} + h_n{}^r h_p{}^s k^{qt} \nabla_m \nabla_r h_{st}] \end{aligned} \tag{124}$$

where, in the first step, we have used the definition of the extrinsic curvature, and, in the second, we have expanded the derivative of the product. Now, each of the first three terms in the last expression in (124) vanishes. The first term vanishes because $h_{[a}{}^m h_{b]}{}^n \nabla_m h_n{}^r = \pi_{[ab]}{}^r = 0$. The second term vanishes for a more subtle reason. Since the indices "m" and "p" of $(\nabla_m h_p{}^s)$ are being projected tangent to S, the index "s" here is normal to S. Thus, for the factor $\nabla_r h_{st}$, the "s" is projected normal to S, and also "t" is being projected normal to S (because of the k^{qt}), But, if both indices "s" and "t" of $\nabla_r h_{st}$ are projected normal to S, we obtain zero. (See the calculation preceding (115).) The third term vanishes by an argument similar to that for the second. Consider the factor $(\nabla_m k^{qt})$. Here, the "q" is being projected normal to S by $(k_q{}^d)$, and the "m" is being projected tangent to S (by $h_a{}^m$). Hence, the "t" in $(\nabla_m k^{qt})$ is tangent to S. But now consider the factor $\nabla_r h_{st}$. Both "s" and "t" are being projected tangent to S, so we get zero. Thus, the right side of (124) becomes simply

$$\begin{aligned} &h_{[a}{}^m h_{b]}{}^n h_c{}^s k^{dt} \nabla_m \nabla_n h_{st} = h_a{}^m h_b{}^n h_c{}^s k^{dt} \nabla_{[m} \nabla_{n]} h_{st} \\ &= h_a{}^m h_b{}^n h_c{}^s k^{dt} \frac{1}{2} \Big[R_{mns}{}^p h_{pt} + R_{mnt}{}^p h_{sp} \Big] \end{aligned}$$

So, we have

$$h_{[a}{}^m h_{b]}{}^n h_c{}^p k_d{}^q \nabla_m \pi_{npq} = -\frac{1}{2} h_a{}^m h_b{}^n h_c{}^p k_d{}^q R_{mnpq} \tag{125}$$

This is the second *Gauss-Codazzi equation*. It looks at the Riemann tensor of M with three indices projected tangent to S and one normal, and expresses it in terms of the rate of change of the extrinsic curvature of S (properly projected). Note that the Riemann tensor of S does not appear above. Thus, curvature of M of a certain sort reflects itself in a derivative of the extrinsic curvature. (Why the projections on the left in (125)? Note that, if any of the h's or the k is replaced by the other, then the result can easily be expressed algebraically in terms of the extrinsic curvature by a single integration by parts.)

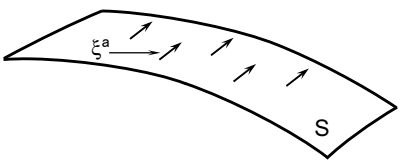
Although the calculations above look rather formidable, time spend staring at them is well worthwhile. It is usually easier to derive the Gauss-Codazzi equations when needed than to go down two flights of stairs to the library to look them up. It should now be clear why we spent so much time at the beginning establishing an index notation!

We shall give examples of the use of these equations shortly.

34. Metric Submanifolds of Co-dimension One

The discussion of the previous sections has been quite general: for a $k-$dimensional submanifold of an $n-$dimensional manifold. Easily the most important case for applications is that in which $k = n - 1$, i.e., the case of co-dimension one. Things can be simplified slightly in this case. We now carry out this simplification.

The crucial consequence of co-dimension one is the following: the space of vectors normal to S at each point of S is one-dimensional. Hence, it consists of multiples of some vector ξ^a. Furthermore, we can normalize ξ^a by $\xi^a\xi_a = \pm 1$. (The choice of sign is forced by the signature of g_{ab} and by S. We shall carry both signs through the calculations.) Thus, the choice of ξ^a is determined uniquely up to sign. (Of course, reversing the sign of ξ^a leaves the sign of $\xi^a\xi_a$ unchanged.) Since the space of normal vectors is one-dimensional, it follows that, if "a" of $\alpha_{ab\ldots c}$ is normal to S, then $\alpha_{ab\ldots c} = \xi^a\beta_{b\ldots c}$. Thus,



$$h_{ab} = g_{ab} \mp \xi_a\xi_b \qquad k_{ab} = \pm\xi_a\xi_b \tag{126}$$

Set

$$\pi_{abc} = \mp\pi_{ab}\xi_c \tag{127}$$

Thus, π_{ab} is symmetric, and both its indices are tangent to S. The extrinsic curvature, a tensor field with three indices, two tangent and one normal to S, reduces to π_{ab}. In co-dimension one, this π_{ab} is called *extrinsic curvature*. To express π_{ab} directly in terms of ξ^a, we proceed as follows:

$$\begin{aligned}\pi_{ab} = -\xi^c\pi_{abc} &= -h_b{}^m\xi^c\, D_a h_{mc} = -h_b{}^m\,[D_a\,(\xi^c h_{mc}) - h_{mc}\,D_a\,\xi^c] \\ &= h_b{}^m\,h_{mc}\,D_a\,\xi^c = h_b{}^m\,D_a\,\xi^m\end{aligned} \tag{128}$$

Eqn. (128) allows an alternative interpretation of the extrinsic curvature in the case of co-dimension one. Let $\tilde{\xi}^a$ be an extension of ξ^a, with $\tilde{\xi}^a\tilde{\xi}_a = \pm 1$, and set $\tilde{h}_{ab} = g_{ab} \mp \tilde{\xi}_a\tilde{\xi}_b$ so $\tilde{h}_{ab}$ is an extension of h_{ab}. Next, note that.

$$\tilde{\xi}^a\,\mathscr{L}_{\tilde{\xi}}\tilde{h}_{ab} = \mathscr{L}_{\tilde{\xi}}(\tilde{h}_{ab}\tilde{\xi}^a) - \tilde{h}_{ab}\,\mathscr{L}_{\tilde{\xi}}\tilde{\xi}^a = 0 - 0 = 0$$

so $\mathscr{L}_{\tilde{\xi}}\tilde{h}_{ab}$ is tangent to S. Thus,

$$\begin{aligned}&\mathscr{L}_{\tilde{\xi}}\tilde{h}_{ab} = \tilde{h}_a{}^m\tilde{h}_b{}^n\,\mathscr{L}_{\tilde{\xi}}\tilde{h}_{mn} = \tilde{h}_a{}^m\tilde{h}_b{}^n\,\mathscr{L}_{\tilde{\xi}}(g_{mn} \mp \tilde{\xi}_m\tilde{\xi}_n) \\ &= \tilde{h}_a{}^m\tilde{h}_b{}^n[\mathscr{L}_{\tilde{\xi}}g_{mn} \mp \tilde{\xi}_m\mathscr{L}_{\tilde{\xi}}\tilde{\xi}_n \mp \tilde{\xi}_n\mathscr{L}_{\tilde{\xi}}\tilde{\xi}_n] = \tilde{h}_a{}^m\tilde{h}_b{}^n\,\mathscr{L}_{\tilde{\xi}}g_{mn} \\ &= \tilde{h}_a{}^m\tilde{h}_b{}^n[\tilde{\xi}^p\nabla_p g_{mn} + g_{mp}\nabla_n\tilde{xi}^p + g_{pn}\nabla_m\tilde{xi}^p] = 2\tilde{h}_a{}^m\tilde{h}_b{}^n\nabla_{(m}\tilde{\xi}_{n)} \\ &= 2\pi_{ab}\end{aligned}$$

The extrinsic curvature thus represents the "Lie derivative of h_{ab} in the direction normal to S". The reason why the extrinsic curvature of the plane and cylinder (in R^s) are what they are is clear from this point of view.

Finally, we substitute (127) into Gauss-Codazzi equations. The resulting (123) is obvious:

$$\mathcal{R}_{abcd} = \pm 2\pi_{c[a}\pi_{b]d} + h_a{}^m h_b{}^n h_c{}^p h_d{}^q R_{mnpq} \tag{129}$$

For (125), note that we might as well contract this equation with ξ^d (for the "d" index is normal to S). Substituting (127) into (125), and nothing that $\xi^q D_m \xi_q = 0$, we have

$$h_{[a}{}^m h_{b]}{}^n h_c{}^p \nabla_m \pi_{np} = \mp \frac{1}{2} h_a{}^m h_b{}^n h_c{}^p \xi^d R_{abcd} \tag{130}$$

These are the *Gauss-Codazzi equations* for co-dimension one.

Example 77. Let $M = R^3$, and let S be the S^2 in M given by $(x^1)^2 + (x^2)^2 + (x^3)^2 = 1$. Let g_{ab} be the metric on M whose components, in the natural chart, are $g_{11} = g_{22} = g_{33} = 1$, others zero. Then S is a metric submanifold. Denote by x_a the vector field on M whose components, in the natural chart, are (x_1, x_2, x_3). Then $\nabla_a x_b = g_{ab}$. On S, x_a is normal to S. From (128)

$$\pi_{ab} = h_b{}^m D_a x_m = h_{ab}$$

We now consider the Gauss-Codazzi equations. Note that R_{abcd}, the Riemann tensor of M, vanishes. Eqn. (130) just reduces to an identity. From (129), we have

$$\mathcal{R}_{abcd} = 2h_{c[a} h_{b]d}$$

This is the expression for the Riemann tensor of the (metric) S^2. Hence,$\mathcal{R} = 2$

Exercise 198. Adapt the example above to find the extrinsic curvature and Riemann tensor for the plane in R^3.

Exercise 199. Adapt the example above to find the extrinsic curvature and Riemann tensor for the cylinder in R^3.

Exercise 200. Adapt the example above to find the extrinsic curvature of S^n in R^{n+1}.

Exercise 201. Let $M = R^4$, with metric, in the natural chart, $g_{11} = g_{22} = g_{33} = 1$, $g_{44} = -1$. others zero. Let S be the submanifold consisting of points with $(x^1)^2 + (x^2)^2 + (x^3)^2 - (x^4)^2 = -1$. (This is called the unit hyperboloid.) Show that S is a metric submanifold, and find its extrinsic curvature and Riemann tensor.

Exercise 202. Show that the Gauss-Codazzi equations are vacuous for submanifolds of dimension one; same for co-dimension zero.

Exercise 203. Show that the result of substituting the right side of (129) into the Bianchi identity on $\mathcal{R}_{abcd}$ is an identity.

Exercise 204. Convince yourself that $h_a{}^m h_c{}^n \xi^b \xi^d R_{mbnd}$ cannot be expressed in terms of the extrinsic and its $D-$derivatives. (Conclusion: There are no additional Gauss-Codazzi equations.)

Exercise 205. Using the interpretation on p. 116, show that, if M is considered as a sub manifold of $M \times R^1$ (a product of manifolds with metric), then the extrinsic curvature of M vanishes.

Exercise 206. using the intuitive interpretation of the Riemann tensor as giving the "change in a vector under parallel transport about an infinitesimal curve", and of the extrinsic curvature as "the date of change of the tangent plane to a submanifold", intuit the validity of the Gauss-Codazzi equations.

35. Orientation

Let M be an $n-$dimensional manifold, and let p be a point of M. We introduce some properties of antisymmetric tensors at p.

Let $\mu_{a\ldots c}$ be an antisymmetric tensor at p of rank n. Then, introducing a chart containing p, we consider the components of $\mu_{a\ldots c}, \tilde{\mu}_{i\ldots j}$. Since $\mu_{a\ldots c}$ is antisymmetric, the only nonzero components are those for which $i \ldots j$ are all different. Hence, $\mu_{a\ldots c}$ is completely determined by the numerical value of the single component $\tilde{\mu}_{1\ldots n}$. Thus, if $\mu_{a\ldots c}$ and $\nu_{a\ldots c}$ are nonzero number α such that $\mu_{a\ldots c} = \alpha\nu_{a\ldots c}$.

Now consider the collection of all nonzero, antisymmetric, rank n, covariant tensors at p. We regard two such tensors as equivalent if the proportionality factor between them is positive. Evidently, there are precisely two equivalence classes. (The vector space of such tensors at p is one-dimensional. We remove the origin, i.e., we exclude the zero tensor. There remain two "pieces", which are precisely the equivalence classes.) A choice of one of these equivalence classes at p is called an *orientation* of the cotangent space at p. Thus, there are precisely two orientations of the cotangent space at p.

The same remarks apply, of course, to contravariant tensors. Similarly, we define an orientation of the tangent space of p. In fact, an orientation of the cotangent space at p defines an orientation of the tangent space at p and vise versa. We shall prove slightly more: if $\epsilon_{a\ldots c}$ is nonzero, rank n, and antisymmetric, then there is precisely one nonzero, rank n, antisymmetric $\epsilon^{a\ldots c}$ (all at p) such that

$$\epsilon^{a\ldots c}\epsilon_{a\ldots c} = n! \tag{131}$$

(The choice of the $n!$ is for convenience. If, with respect to a chart, $\epsilon_{1\ldots n} = 1$. then, with respect to that same chart, $\epsilon^{1\ldots n} = 1$). To prove this, let $\mu^{a\ldots c}$ be nonzero, and set $\alpha = \mu^{a\ldots c}\epsilon_{a\ldots c}$. Note (e.g., using a chart) that $\alpha \neq 0$. Hence, $\epsilon^{a\ldots c} = n!\,\alpha^{-1}\mu^{a\ldots c}$ is the (obviously unique) solution of (131). The $\epsilon^{a\ldots c}$ of (131) is called the *inverse* of $\epsilon_{a\ldots c}$. Noting that, if two nonzero $\epsilon_{a\ldots c}$'s differ by a positive factor, then so do their inverses, we obtain a correspondence between the (two) orientations of the cotangent space and those of the tangent space.

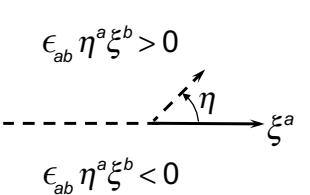
The sense in which $\epsilon^{a\ldots c}$ is the "inverse" of $\epsilon_{a\ldots c}$ is the following. Let $\mu_{a\ldots c}$ be antisymmetric. Then, since μ is a multiple of $\epsilon_{a\cdots c}$, the left side of

$$\frac{1}{n!}\mu_{m\cdots n}\epsilon^{m\cdots n}\epsilon_{a\ldots b} = \mu_{a\cdots c} \tag{132}$$

is a manifold of the right. Contracting with $\epsilon^{a\ldots c}$, we see that (132) holds.

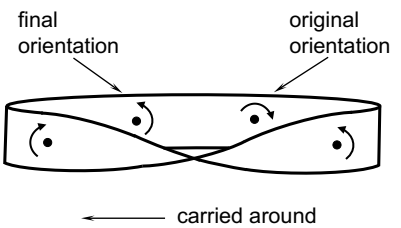
Roughly speaking, an orientation of a manifold M consists of a continuous assignment of an orientation to the tangent space of each point of M. More precisely, consider two tensor fields, $\mu_{a\ldots c}$ and $\nu_{a\ldots c}$, on M, which are antisymmetric and of rank n. Suppose neither vanishes anywhere. Hence, $\mu_{a\ldots c} = \alpha\nu_{a\ldots c}$, where α is a scalar field on M. (Exercise 207. Prove smooth.) These two fields will be regarded as equivalent if α is everywhere positive. (Note that, in any case, α vanishes nowhere.) An equivalence class of such fields defines an *orientation* of M. As we shall see in a moment, a manifold M may have no orientation. Suppose that M is connected. Then any scalar field on M which vanishes nowhere is either everywhere positive or everywhere negative. (Exercise 208. Prove this statement.) In this case, therefore, if there is any orientation on M, there are precisely two.

Finally, we give a geometrical picture of orientation in the case of dimension two. Let ϵ_{ab} be a nonzero skew tensor at a point p of a $2-$dimensional manifold. Fix a vector ξ^a at p, and consider the quantity $\epsilon_{ab}\eta^a\xi^b$ as η^a



varies through vectors at p. This quantity is zero for $\eta^a = \xi^a$, and, as η^a "rotates" away from ξ^a, becomes either positive or negative. It is again zero for $\eta^a = -\xi^a$. Hence, the direction in which η^a must rotate from ξ^a to make the quantity above positive defines a "direction of rotation" from ξ^a. Thus, an orientation on a 2−dimensional manifold M represents a "direction of rotation" in the tangent space at each point of M.

The question of whether a manifold has an orientation is the question of whether or not there exists a continues choice of rotation direction at each point. It is perhaps not surprising that a Mobius strip has no orientations.



36. Integrals

We now introduce the notion of an integral over a manifold. We begin with integrals in R^n. Let U be an open subset of R^n, and let $f(x)$ (where $x = (x^1, \ldots, x^n)$) be a function on U. Then, of course, the Riemann integral,

$$\int_U f(x)\, dx^1 \cdots dx^n \tag{133}$$

is well defined. (When we write an integral, the assumption that the integral converges absolutely is automatically in force. if an integral does not converge, it does not converge, and the discussion ends.) Suppose we now perform a coordinate transformation, $x'^1 = x'^1(x), \ldots, x'^n(x) = x'^n(x)$. Then, from elementary calculus,

$$\int_U f(x)\, dx^1 \cdots dx^n = \int_U f(x(x')) \left|\frac{\partial x}{\partial x'}\right| dx'^1 \cdots dx'^n \tag{134}$$

where $|\frac{\partial x}{\partial x'}|$ is the Jacobian, i.e., the absolute value of the determinant of the $n \times n$ matrix $\partial x^i / \partial x'^j$ $(i, j = 1, \ldots, n)$.

It is clear from (134) at there is no immediate definition of the integral of a function f over an open region U of a manifold M. The most obvious definition would be as follows: choose a chart containing U, and define $\int_U f$ to be (133). But this is no good, for, by (134), the result depends on the choice of chart. Of course, any reasonable definition of an integral depends only on the quantity being integrated and the region of M over which the integration takes place.

How do we arrange things so we get a reasonable notion of integration in a manifold? Let $f_{a\ldots c}$ be an $n-$form on the ($n-$dimensional) manifold M, and let U be an open subset of M sufficiently small that there exists a chart on U M containing U. Choose two such charts, with coordinates x^i and x'^i $(i = 1, \ldots, n)$. Then the components of $f_{a\ldots c}$ in these charts are related by

$$f'_{i\cdots j} = \Sigma f_{k\cdots l} \frac{\partial x^k}{\partial x'^i} \cdots \frac{\partial x^k}{\partial x'^j} \tag{135}$$

From (135), it follows that

$$f'_{i\cdots j} = f_{i\cdots j} \det\left(\frac{\partial x^k}{\partial x'^l}\right) \tag{136}$$

(Proof: The left side of (136) is certainly proportional to the right, by Sect. 35. To prove equality, contract both sides with $\epsilon^{i\cdots j}$, use (132), and the fact that

$$\det K_a{}^b = \frac{1}{n!} K_a{}^{b\cdots} K_d{}^c \epsilon^{a\cdots d} \epsilon_{b\cdots c} \tag{137}$$

for a tensor $K_a{}^b$ where $\epsilon^{a\ldots d}$ and $\epsilon_{b\ldots c}$ are inverses.)

We are now almost prepared to define integration on a manifold. The only remaining difficulty is the presence of an absolute value (in the Jacobian) in (134). To correct this, we only admit coordinates x^i $(i = 1, \ldots, n)$ in which $\epsilon_{1\cdots n} > 0$, where $\epsilon_{a\cdots b}$ represents an orientation on M. Let $f_{a\ldots c}$ be an $n-$form on the oriented ($n-$dimensional) manifold M. Choose a chart compatible (in the sense above) with the orientation of M, and define the left side of

$$\int_U f_{a\cdots c} ds^{a\cdots c} = n! \int_U f_{1\cdots n}\, dx^1 \cdots dx^n \tag{138}$$

by the right side. The result is independent of choice of chart, for, by (134) and (136),

$$\int_U f_{1\cdots n} dx^1 \cdots dx^n = \int_U f_{1\cdots n} \left|\frac{\partial x}{\partial x'}\right| dx'_1 \cdots dx'^n = \int_U f'_{1\cdots n}\, dx'^1 \cdots dx'^n$$

Thus, integration should be performed on an oriented manifold, and the proper thing to integrate is an $n-$form. (These are not all that much different from functions. Each forms are one-dimensional vector space at each point. The nice thing about forms is that their components have the correct behavior, e.g., (134), under coordinate transformations.

As in elementary calculus, we define integrals (of $n-$forms on an oriented manifold) over arbitrary open regions U by the requirement that, if U is contained in a chart, the integral is given by (138), and for U and V disjoint,

$$\int_U + \int_V = \int_{U\cup V} \tag{139}$$

Finally, note that, if $\mu_{a\cdots c}$ and $\nu_{a\cdots c}$ are $n-$forms, s a constant, then

$$\int_U (\mu_{a\cdots c} + s\nu_{a\cdots c}) ds^{a\ldots c} = \int_U \mu_{a\cdots c} ds^{a\cdots c} + s\int_U \nu_{a\cdots c} ds^{a\cdots c} \tag{140}$$

Integration on a a manifold is not much different from integration in elementary calculus. The only difference is that one integrates an $n-$form instead of a function.

The thing one integrates over a $k$dimensional submanifold of M is a $k$form. Let $\psi : S \to M$ be a submanifold of M, and suppose that S is oriented (i.e., at each point of $\psi[S]$ an equivalence class of antisymmetric, rank k tensors, all of whose indices are tangent to S, is specified). The left side of

$$\int_{\psi[U]} \omega_{a\cdots b}\, ds^{a\cdots b} = \int_U (\underset{\leftarrow}{\psi}\, \omega_{a\cdots b})\, ds^{a\cdots b} \tag{141}$$

is defined by the right, where U is an open subset of S, and $\omega_{a\cdots b}$ is a $k-$form on M.

37. Stokes' Theorem

Recall the equation of the fundamental theorem of calculus

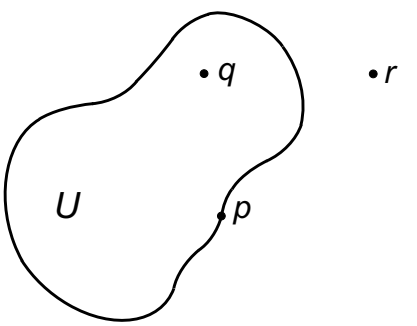
$$\int_a^b f'(x)\,dx = f(b) - f(a) \tag{142}$$

There is a generalization of this result to manifolds. The integral over the real line in (142) is replaced by an integral over a manifold, the function f in (142) by a form, and the derivative in (142) by the exterior derivative. Making these replacements, we have

$$\int_U D(f)_{ma\cdots b}\,ds^{ma\cdots b} = \int_{\partial U} f_{a\cdots b}\,ds^{a\cdots b} \tag{143}$$

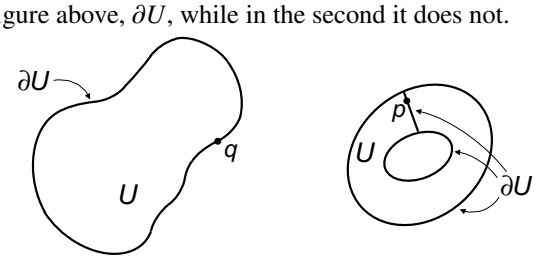
where $f_{a\cdots b}$ is an $(n-1)-$form, U is an open subset of M, and ∂U is the "boundary" of U. Accompanied by an appropriate collection of definitions and qualifications, Eqn. (143) is indeed true. It is this result we wish to establish in the present section.

Let M be an $n-$dimensional manifold, U and open subset of M. The *boundary* of U, ∂U, is the collection of all points p of M such that p is not in U, but every chart containing p intersects U. This definition corresponds, of course, to the intuitive notion of the boundary of an open set. (In the figure on the right, p is in the boundary of U, but neither q nor r are in ∂U.)



In order that the right side of (143) make sense, it is necessary that ∂U be an $(n-1)-$dimensional submanifold of M.

We next deal with the orientation in (143). Let M be oriented. Thus, at each point of M, we have selected one of the two classes of equivalent (nonzero) $n-$forms. Now the integral on the left in (143) makes sense. What orientation should we choose for ∂U on the right? Fix a point p of ∂U. Then the covariant vectors (in M) at p which are normal to the $(n-1)-$dimensional sub manifold M form a one-dimensional subspace of the cotangent space of p. This one-dimensional vector space is divided into two pieces by its zero element. We want to regard one of these pieces as consisting of "outgoing normals" (i.e., heading away from U), and the other as consisting of "ingoing normals". Such a distinction would, evidently, be impossible of U "met ∂U on both sides of ∂U". (E.g., as in p in the figure below on the right rather than q.) We are thus led to the following definition: ∂U will be said to *border* the open set U if ∂U is an $(n-1)-$dimensional submanifold and if, for each point p of ∂U, there is a chart containing p such that the intersection of this chart with U consists precisely of the points of the chart with $x^n < 0$. Thus, in the first figure above, ∂U, while in the second it does not.



A covariant vector μ_a normal to ∂U at a point p of ∂U will be said to be an *outgoing* normal if, with respect to the chart above, the $n^t h$ component of μ_a is positive. Evidently, ν_a is another outgoing normal if and only if ν_a is a positive multiple of μ_a.

Finally, we are ready to induce an orientation on ∂U from that of M. Let p be a point of ∂U. and let $\lambda^{a\cdots c}$ be antisymmetric of rank n, and in the equivalence class defining the orientation of M. Let μ_a be an outgoing normal to ∂U. Then $\mu_a \lambda^{ab\cdots c}$ is antisymmetric, of rank $(n-1)$, and with all indices tangent to ∂U. This tensor thus defines an orientation of ∂U. We call this orientation the induced orientation on ∂U. Note that, if we reverse the orientation originally assigned to M, we also reverse the induced orientation on ∂U.

Thus, if we assume that ∂U borders U and use the induced orientation on the right in (143), everything in that equation is well-defined. Is the equation now true in general? Unfortunately, the answer is no. We still require one further condition. In order to see why this condition is necessary, we begin the proof of (143).

Let U be the region of R^n consisting of points $(x^1, \ldots, x^n)$ with $0 < x^1 < 1, \ldots, 0 < x^n < 1$. Let $f(x)$ be a function on R^n. Then

$$\int_U \frac{\partial f}{\partial x^1}\, dx^1 \cdots dx^n = \int_{x^1=1} f\, dx^2 \cdots dx^n - \int_{x^1=0} f\, dx^2 \cdots dx^n \tag{144}$$

where the two integrals on the right are surface integrals over appropriate faces on the cube U. Next, let $f_{i\ldots j}(x)$ be the components of some $(n-1)-$form on R^n. Then the components of the exterior derivative of this form are

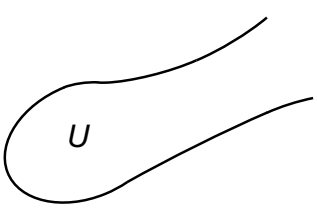
$$\frac{1}{n}\Big[\frac{\partial}{\partial x^1} f_{2\cdots n} + \epsilon \frac{\partial}{\partial x^2}\, f_{3\cdots n1} + \cdots + \epsilon \frac{\partial}{\partial x^n}\, f_{1\cdots(n-1)}\Big] = D(f)_{1\cdots n} \tag{145}$$

$$\epsilon = \begin{cases} 1 & n \text{ odd} \\ -1 & n \text{ even} \end{cases}$$

From (144) and (145), we see that Eqn. (143) is valid for this U provided we take for ∂U on the right the union of the (plane) faces of U (leaving off the corners and edges, etc. We know from elementary calculus that these are irrelevant. Strictly speaking, (143) is not meaningful for our U because the actual ∂U is not a sub manifold: it has corners.)

From these remarks, the following statement should be clear: Eqn. (143) is valid if there is a chart containing U and ∂U such that the image of $U_\cup \partial U$ in R^n is closed and bounded.

The additional condition we require is that U does not "go off to infinity in M" so that "some of the surface integral on the right in (143) is not included in (143) because the appropriate boundary is off at infinity". In other words, we must ensure that none of the flux escapes to infinity: that all is properly taken into account by the surface integral on the right in (143). For the U on the right, for example, we could not guarantee that (143) is valid.



A subset C of M will be said to be chart-compact if there is a chart containing C such that the image of C in R^n is closed and bounded. A subset C of M is said to be *compact* if C can be written as a union of chart-compact sets. (As usual, we have merely defined a standard topological notion on M, but using chart instead of topological notion.)

The following statement is true:

Theorem 22. Let U be an open subset of M such that ∂U borders U, and such that $U_\cup \partial U$ is compact. Let M be oriented. Then, using the induced orientation on ∂U, Eqn. (143) holds. We have not, of course, proven Theorem 22. It should be clear from the discussion above, however, that the techniques of elementary calculus together with a certain amount of work will suffice to write out a full proof.

Since integrals over sub manifold reduce to integrals over manifolds, Theorem 22 is also true for submanifolds. These results are called *Stokes' Theorem*. Essentially every situation in practice in which an integral "cancels" a derivative is a special case of Stokes' theorem.

Example 78. Let S be a one-dimensional submanifold of M, so S consists of a curve. Let p and q consist of the endpoints of this curve. Then Stokes' theorem becomes

$$\int_S D(\alpha)_a \, ds^a = \alpha(q) - \alpha(p)$$

for any scalar field α.

Example 79. Let U consist of the points of R^2 with $(x^1)^2 + (x^2)^2 < 1$. Then ∂U consists of points with $(x^1)^2 = (x^2)^2 = 1$. In this case, ∂U borders U.

Example 80. Let U consist of points of R^2 with $(x^1)^2 + (x^2)^2 \neq 1$. Then ∂U is as above, but now ∂U does not border U.

Exercise 207. Find the orientations of S^2, of R^2.

Exercise 208. Using Example 78, prove that a scalar field on a connected manifold, having vanishing gradient, is constant.

Exercise 209. Use Theorem 22 to prove that the exterior derivative, applied twice in succession, gives zero. (Hint: the left side of (143) remains the same if U is replaced by any other open set with the same boundary as U.)

Exercise 210. Verify (143) explicitly in R^2 with the U of Example 79, and with f_a with components $(x^2, -x^1)$. That R^2 is not compact.

Exercise 211. Show that the manifold S^2 itself is compact. That R^2 is not compact.

Exercise 212. Consider the geometrical interpretation of orientation (for a two-dimensional manifold) in Sect. 35. Obtain a similar interpretation for a one-dimensional submanifold. Using these interpretations, describe the operation of obtaining the induced orientation on the boundary of an open subset of a two-dimensional manifold.

Exercise 213. Derive (142) from (143).

38. Integrals: The Metric Case

On a manifold with metric, there is an alternative way of writing integrals. This formation is often the most convenient and suggestive.

Let M be an oriented $n-$dimensional manifold with metric g_{ab}. Let $\omega_{a\cdots c}$ be an $n-$form at a point p of M, where this $n-$form is consistent with the orientation of M. Then, since g_{ab} is invertible, the quantity

$$\omega_{a\cdots c}\omega^{a\cdots c} = \omega_{a\cdots c}(g^{am}\cdots g^{cn}\,\omega_{m\cdots n}) \tag{146}$$

is nonzero. If $\omega_{a\cdots c}$ in (146) is replaced by $\alpha\omega_{a\cdots c}$, where α is a number, then the expression (146) is multiplied by α^2. Thus, there exists precisely one tensor $\epsilon_{a\cdots c}$ which is antisymmetric in all indices, of rank n consistent with the orientation of M, and satisfies

$$\epsilon_{a\cdots c}\epsilon^{a\cdots c} = \pm n! \tag{147}$$

(The plus sign must be used in (147) if the signature of g_{ab} contains an even number of $-$'s; otherwise the minus sign.) The tensor field $\epsilon_{a\cdots c}$ is called the *alternating tensor* associated with g_{ab} and the orientation of M. If the orientation of M is reversed, then the sign of $\epsilon_{a\cdots c}$ is reversed.

Let $\omega_{a\cdots c}$ be an $n-$form on M. Then

$$\omega_{a\cdots c} = \pm\frac{1}{n!}\,\epsilon_{a\cdots c}\,(\epsilon^{m\cdots n}\omega_{m\cdots n}) \tag{148}$$

Thus, the $n-$form is completely and uniquely determined by the value of the scalar field $\omega_{a\cdots c}\epsilon^{a\cdots c}$. In other words, on an oriented manifold with metric, there is a one-to-one correspondence between $n-$forms and scalar fields. We define the left side of

$$\int_U f\,dS \;=\; \int_U \frac{1}{n!}\,f\,\epsilon_{a\cdots c}\,dS^{a\cdots c} \tag{149}$$

by the right. We can integrate scalar fields over open regions of M. In particular,

$$\int_U 1\,dS \;=\; \int_U \frac{1}{n!}\,\epsilon_{a\cdots c}\,dS^{a\cdots c} \tag{150}$$

is called the *volume* of U. Thus, a metric gives rise, in particular, to the notion of volume.

Next, let $\mu_{a\cdots b}$ be an $(n-1)-$form on M.Then, setting

$$\nu^a = \epsilon^{ab\cdots c}\mu_{b\cdots c} \tag{151}$$

we have

$$\mu_{m\cdots c} = \nu^a\epsilon_{ab\cdots c}\,(\pm\frac{1}{(n-1)!}) \tag{152}$$

Thus, we obtain a natural, one-to-one correspondence between $(n-1)-$forms and vector fields on M. If S is an oriented sub manifold of M, of co-dimension one, we define the left side of

$$\int_S \nu^a\,dS_a = \frac{1}{(n-1)!}\;\int_S (\nu^a\,\epsilon_{ab\cdots c})\,dS^{b\cdots c} \tag{153}$$

by the right side. In particular, if S is a metric submanifold we define the *area* of S by

$$\int_S \nu^a\,dS_a \tag{154}$$

where ν^a is a unit normal vector.

We can now express n-form in terms of scalar fields, and $(n-1)$ forms in terms of vector fields. Hence, the exterior derivative (which takes an $(n-1)-$form to an $n-$form) should take a vector field to a scalar field. What is the expression for this action? Let $\omega_{a\cdots b}$ be an $n-1$ form, and v^a its corresponding vector field. Then, from the concomitant (54), we have

$$D(\omega)_{ma\cdots b}\,\epsilon^{ma\cdots b} = \nabla_m\,(\epsilon^{ma\cdots b}\,\omega_{a\cdots b}) = \nabla_m v^m \tag{155}$$

where ∇_m is the derivative operator defined by g_{ab}. Thus, the operation is simply that of taking the divergence using the derivative operator defined by g_{ab}.

Finally, substituting (155) into (143), we have

$$\int_U (\nabla_a v^a)\,dS \;=\; \int_{\partial U} v^a\,dS_a \tag{156}$$

This, of course, is the familiar form for Gauss' law.

Example 81. Let M be an oriented, three-dimensional manifold with metric. Then Theorem 22, for a $3-$dimensional submanifold, is Gauss' law; for a $2-$dimensional sub manifold is what is usually called Stokes' theorem; and for a $1-$dimensional sub manifold is the line integral theorem (Example 78). Thus, the three standard integral theorems for a three-dimensional manifold are precisely the three versions of Stokes' theorem.

Exercise 214. Prove that $\nabla_a \epsilon_{b\cdots c} = 0$. where $\epsilon_{b\cdots c}$ is the alternative tensor, and ∇_a is the derivative operator defined by the metric. (Hint: Take a derivative of (147).)

Exercise 215. Find the volume of S^2 (with the usual metric).

Exercise 216. Find the area of S^2 consider as a submanifold of R^3 (with usual embedding and metric).

Exercise 217. Let M be an oriented $n-$dimensional manifold with metric. Extend the discussion above to obtain a correspondence between $k-$forms on M and $(n-k)-$forms. Express the action of the exterior derivative.

Appendix

Math 279 **Test** February 3, 1972
Due: February 19, 1972

The twelve problems below are arranged approximately is order of difficulty, from easiest to hardest. Do the hardest six you can.

1. Let ξ^a be a nonzero vector at a point p of the $n-$dimensional manifold M, and let $(\tilde{\xi}^1 \ldots, \tilde{\xi}^n)$ be n numbers, not all zero. Prove that there exists a chart with respect to which the components of ξ^a are $(\tilde{\xi}^1, \ldots, \tilde{\xi}^n)$.

2. Let $R_{abc}{}^d$ be a tensor at the point p of the paracompact manifold M, satisfying $R_{abc}{}^d = R_{[ab]c}{}^d$ and $R_{[abc]}{}^d = 0$. Prove that there exists a derivative operator on M whose Riemann tensor at p is this $R_{abc}{}^d$.

3. Find two distinct metrics on R^2 which define the same derivative operator. Find two distinct derivative operators on R^2 which have the same Riemann tensor.

4. Let η^a be a vector field on the manifold M such that $\mathscr{L}_\xi \eta^a = 0$ for every ξ^a. Prove that $eta^a = 0$

5. Prove Eqn. (31).

6. Let $\psi : M \to M'$ be smooth, and let U be an open subset of M'. Prove that $\psi^{-1}[U]$ is an open subset of M.

7. Let M and M' be manifolds. Find a natural diffeomorphism between $T(M \times M']$ and $TM \times TM'$.

8. Let M be a three-dimensional manifold with metric g_{ab}. Prove that, if the Ricci tensor vanishes, then so does the Riemann tensor.

9. Find a smooth, onto mapping $\psi : M \to M'$ such that $\vec{\psi}_p$ is an isomorphism for each point p of M, but such that ψ is not a diffeomorphism.

10. Find a connected manifold M such that, for any point p of M, $M - p$ is diffeomorphic with M.

11. develop an algebra of the concomitant (58) along the lines of Example 30.

12. For two totaly symmetric contravariant tensors $\alpha^{a_1 \ldots a_p}$ and $\beta^{a_1 \ldots a_q}$ at a point, set $\alpha \cap \beta = \alpha^{a_1 \ldots, a_p} \beta^{a_{p+1} \ldots a_{p+q}}$. A symmetric contravariant tensor $\gamma^{a \ldots, c}$ at this point will be said to be *prime* if it cannot be written as $\alpha \cap \beta$ with α and β of nonzero rank. Prove that every symmetric contravariant tensor at the point can be decomposed into a product of primes, and that this decomposition is unique except for order and scalar factors.

Math 279 **Final Examination** February 29, 1972
Due: March 9, 1972

The twenty-four problems below are arranged into eight groups of three. The problems within each group are arranged in order of difficulty. Do the hardest ten you can, including at least one problem from each of the eight groups.

Group A.

1. Let μ_{ab} be an antisymmetric tensor at a point, and ν_a nonzero vector. Suppose $\mu_{[ab}\nu_{c]} = 0$. Prove that there exists a vector ω_b at the point such that $\mu_{ab} = \nu_{[a}\omega_{b]}$.

2. Let ϵ^{abcd} be the alternating tensor on an oriented four-dimensional manifold with positive-definite metric. Prove that $\epsilon^{abmn}\epsilon_{cdmn} = \pm\frac{1}{4}\delta^{[ac}\delta^{b]}{}_d$.

3. Let K^{abc} be a tensor at a point. Prove that K^{abc} can be written uniquely in the form $K^{abc} = \alpha^{abc} + \beta^{abc} + \gamma^{abc} + \delta^{abc}$ with $\alpha^{abc} = \alpha^{(abc)}, \beta^{abc} = \beta^{[abc]}, \gamma^{abc} = \gamma^{(ab)c}, \gamma^{(abc)} = 0$, $\delta^{abc} = \delta^{[ab]c}$, and $\delta^{[abc]} = 0$.

Group B.

4. Let p and q be distinct points of S^3. Prove that $S^- p - q$ is diffeomorphic with $S^2 \times R$.

5. Let M and M' be $n-$dimensional manifolds with vector fields ξ^a and ξ'^a, respectively, each of which vanishes nowhere. Let p be a point of M, and p' be a point of M'. Prove that there exists an open set U containing p, and U' containing p', such that there exists a diffeomorphism from U to U' which takes ξ^a to ξ'^a.

6. Find manifolds M and M' which are not diffeomorphic, such that $M \times R$ is diffeomorphic with $M' \times R$.

Group C.

7. Prove that the interior derivative of an $n-$form on an $n-$dimensional manifold always vanishes.

8. Let ξ^a and η^a be vector fields, with η^a vanishing nowhere. If $\mathscr{L}_\xi(\eta^a \eta^b) = 0$, prove that $\mathscr{L}_\xi \eta^a = 0$.

9. Let $\epsilon_{a\ldots c}$, be an $n-$form, vanishing nowhere, on an $n-$dimensional manifold. Let $\epsilon^{a\ldots c}$ be its inverse. Prove that $\mathscr{L}_\xi \epsilon_{a\ldots c} = 0$ if and only if $\mathscr{L}_\xi \epsilon^{a\ldots c} = 0$.

Group D.

10. On a manifold (of dimension greater than two) with metric, let $R_{abcd} = C\, g_{a[c} g_{d]b}$. Prove that $C = const$.

11. On a manifold with metric, a vector field ξ^a is said to define a *constant of the motion* if, for any geodesic with tangent η^a, $\xi_a \eta^a$ is constant along the geodesic. Prove that ξ^a defines a constant of the motion if and only if ξ^a is a Killing vector.

12. Prove that, on a two-dimensional manifold with metric, there exists a scalar field α such that $R_{abcd} = \alpha g_{a[c}\, g_{d]b}$.

Group E.

13. Let $\psi : \; M \to M'$ be smooth. Prove that the subset of $M \times M'$ consisting of points of the form $(p, \psi(p))$ is a submanifold.

14. Let $\psi : \; M \to M'$ and $\Phi : \; N \to N'$ be smooth. Prove that $\Lambda : M \times N \to M' \times N'$ defined by $\Lambda(p, q) = (\psi(p), \Phi(q))$ is smooth.

15. Let $psi : \; M \to M'$ be smooth. When is $\underleftarrow{\psi}$ a one-to-one correspondence between covariant vector field on M' and those on M?

Group F.

16. Prove that the image of a cross-section of a vector bundle is a submanifold of the bundle space.

17. Prove that two derivative operators on a manifold are the same if and only if they define the same horizontal subspace of each tensor bundle.

18. Let M be a manifold with positive-definite metric. Find a natural positive-definite metric on the tangent bundle of M.

Group G.

19. Let ξ^a be a complete vector field on M, and p a point of M. Prove that ξ^a is complete on $M - p$ if and only if $\xi^a = 0$ at p.

20. Classify, the two-dimensional Lie algebras.

21. find two complete vector fields on a manifold whose sum is not complete.

Group H.

22. Let S (a subset of a manifold M) be a submanifold, and let T (a subset of S) be a submanifold of S. Prove that T is a submanifold of M.

23. Let S be the subset of R^3 consisting of points with $(x^1)^2 - (x^2)^2 - (x^3)^2 = 1$. Let g_{ab} be the metric on R^3 with components $g_{11} = -1$, $g_{22} = g_{33} = 1$, others zero. Find the extrinsic curvature and intrinsic metric of S.

24. Find a one-form ω_a on S^1 R such that $D(\omega)_{ab} = 0$, but such that, for some closed curve C, $\int_C \omega_a \, dS^a \neq 0$.

Robert Geroch

Infinite-Dimensional Manifolds

1975 Lecture Notes

With 65 Figures

MINKOWSKI
Institute Press

Robert Geroch
Enrico Fermi Institute
University of Chicago

© Robert Geroch 2013
All rights reserved. Published 2013

ISBN: 978-1-927763-15-5 (softcover)
ISBN: 978-1-927763-16-2 (ebook)

Minkowski Institute Press
Montreal, Quebec, Canada
http://minkowskiinstitute.org/mip/

For information on all Minkowski Institute Press publications
visit our website at http://minkowskiinstitute.org/mip/books/

Contents

1 Introduction 1

2 Banach Spaces 2

3 Banach Spaces from Banach Spaces 6

4 Open Mapping Theorem 8

5 Splitting 10

6 Bounded Linear Mappings 14

7 Derivatives 17

8 Mean Value Theorem; Inverse Mapping Theorem 21

9 Manifolds 24

10 Mappings of Manifolds 29

11 Scalar Fields 31

12 Vector Fields 34

13 Tensor Products 39

14 Tensor Spaces 43

15 Natural Tensors 45

16 Natural Tensors: Continued 50

17 Tensor Fields 53

18 Tensor Bundles 55

19 Lie Derivatives 57

20 Integral Curves 62

21 Geometry of Lie Derivatives 66

22 Exterior Derivatives 69

23 Derivative Operators 73

24 Derivatives: Summary 75

25 Partial Differential Equations 76

26 Riemannian Geometry 80

Appendix 83

1. Introduction

A manifold is, in general terms, "a space which, locally, looks like some simple space, although it may, in the large, look quite different from that simple space". Different choices of this "simple space" lead to different types of manifolds. By far the most common such choice is $\mathbb{R}^n$ (the finite-dimensional vector space consisting of $n-$tuples of real numbers). In this case, the corresponding manifolds are said to be finite-dimensional (specifically, $n-$dimensional), and the study of such manifolds and their structure is called differential geometry.

Finite-dimensional manifolds are important in both mathematics and physics. In mathematics, the study of finite-dimensional manifolds, differential geometry, is an end in itself. However, there exist in addition applications of manifolds to other areas of mathematics. An outstanding example is that of ordinary differential equations. The question of the existence, uniqueness, and structure of the solutions of ordinary differential equations can be reduced to the study of vector fields and their integral curves on manifolds. In this way, one recasts the subject of ordinary differential equations into an elegant, geometrical, and remarkably simple form. In physics, for example, the space-time of general relativity is found to have the structure of a finite-dimensional (in fact, four-dimensional) manifold. As a second example from physics, the space of configurations of a mechanical system becomes a manifold (where the dimension is what is called the number of degrees of freedom of the system).

One can also introduce manifolds based on spaces which are "larger than finite-dimensional vector spaces", i.e., on "infinite-dimensional spaces". Although it is perhaps true that the possible applications of such manifolds have not yet been exploited fully, there are already clear indications that these applications will be rich and far-ranging. One may expect, for example, that the subject of partial differential equations (particularly, hyperbolic and parabolic equations) can be formulated in the same geometrical language as was possible for ordinary differential equations. In physics, infinitive-dimensional manifolds have already found their way into such areas as general relativity and quantum field theory.

One might, therefore, wish to learn about infinite-dimensional manifolds. The next issue one must resolve is this: On what "simple (but infinite-dimensional) spaces" should these manifolds be based? There is only one such choice for which, as far as I am aware, the subject has been worked out in any detail, namely, that in which the simple space is what is called a Banach space. It is my own view that manifolds based on Banach spaces might be "too right in structure" for certain applications. That is to say, although one obtains a large number of strong theorems in this case, it may sometimes turn out to be the case that it is not possible to make the space one is considering in some application into a manifold based on a Banach space. Nonetheless, one has to begin this subject somewhere, and it seems advisable to begin in that regime in which i) many theorems are available, and ii) the subject has been worked out in detail. If one should later decide that other types of manifolds will be more fruitful, then one will at least have a feeling for the kinds of things which are and are not likely to be true, the techniques which are likely to lead to proofs, and where to look for counterexamples.

We shall here study manifolds based on Banach spaces: their structure, and the kinds of objects which can be placed on them.

2. Banach Spaces

In this section, we define a Banach space, give a few examples and introduce related notions.

A *Banach space* consists of two things:

1. A real vector space E. That is to say, E is a set (whose elements are called *vectors*), together with a rule which assigns to any two vectors another (called their *sum*), and another rule which assigns to any vector and any real number a vector (called the *product* of the number and the vector), subject to the usual conditions for a vector space (namely, addition is commutative and associative, there is an additive identity, there are additive inverses, multiplication of vectors by numbers is distributive, and $1 \cdot x = x$ for any vector x).

2. A norm on the vector space E. That is to say, we are given a mapping from the vector space E to the reals. The real number which is the image of the vector x under this mapping is written $|x|$, and is called the *norm* of x. This mapping must satisfy the following three conditions: i) for any vector x, $|x| \geq 0$, with equality if and only if $x = 0$, ii) for any two vectors x and y, $|x + y| \leq |x| + |y|$, and iii) for any vector x and any number a, $|ax| = |a|\,|x|$. [These three conditions are perhaps rather natural if one interprets the norm of a vector as its "length".]

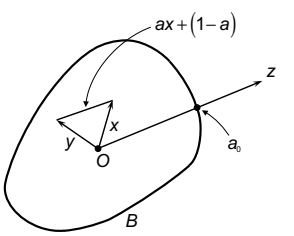
Finally, in order that E, $|\ |$ be a Banach space, it is necessary that it be complete, as described below.

Let E be a real vector space, with norm $|\ |$. Let $x_1, x_2, \ldots$ be a sequence of vectors in E. This sequence is called a *Cauchy sequence* if the following property is satisfied: given any positive number ϵ, there exists an integer N such that $|x_i - x_j| \leq \epsilon$ whenever $i \geq N$ and $j \geq N$. [In intuitive terms, the elements of a Cauchy sequence "get closer and closer to each other along the sequence." This sequence is said to *converge* to vector x if the limit of $|x - x_j|$ as j goes to infinity is zero (i.e., in intuitive terms, if "the elements of the sequence get closer and closer to x"). It is easily checked that a sequence that converges to some x is automatically a Cauchy sequence. Our E, $|\ |$ is said to be *complete* if the converse also holds therein, i.e., if every Cauchy sequence in E, $|\ |$ converges to some vector in E.

Thus, a Banach space is a complete normed vector space. We now give some examples.

Example. Let E be the vector space of real numbers (so addition of vectors is addition of real numbers; and multiplication of vectors by numbers is multiplication of real numbers). Let the norm of vector (real number) x be the absolute value. The three properties of the norm are elementary facts about the absolute value; completeness is a basic fact about the real number system. Thus, we obtain a Banach space.

Example. Let E be any finite-dimensional vector space. Choose any subset B of E which has the following two properties: i) B is convex, i.e., given any vectors x and y in B, and any number a with $0 \leq a \leq 1$, the vector $ax + (1 - a)y$ is in B, and ii) B is radial, i.e., given any non-zero vector x, there is a number $a_0 > 0$ such that ax is in B whenever $0 < a < a_0$ and ax is not in B whenever $a > a_0$. [The first property requires that "the line segment joining any two vectors in B lies entirely within B"; the second requires that "any ray emanating from 0 is within B for a while, and then leaves and remains outside of B".] We use this B to define a norm on E as follows; for x any nonzero vector in E, set $|x| = 1/a_0$, where a_0 is the number in property ii) above. Then the first and third properties for a norm follow from the fact that B is radial, while the second follows from the fact that B is convex. Completeness is



easily checked. Thus, we obtain a Banach space. [For example, if B is the "unit sphere" about 0, then the norm of a vector is just its "Euclidean length".]

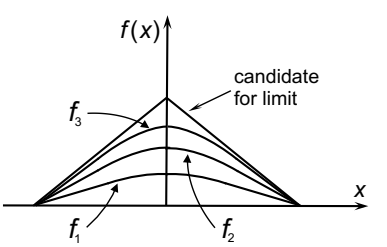
Indeed, it is not difficult to show that this last example yields the "generic" finite-dimensional Banach space, in the sense that every such Banach space arises as in the example. [Sketch of proof: Given a finite-dimensional space, E, $|\ |$, denote by B the set of all vectors x with $|x| < 1$. This set B satisfies the two conditions above, and generates, by the construction above, the original norm on E.]

Example. Denote by E the collection of all sequences of real numbers, $(r_1, r_2, \ldots)$, which are bounded. Define addition and multiplication by numbers component-wise (*i.e.*, $(r_1, r_2, \ldots) + a(r'_1, r'_2, \ldots) = (r_1 + a\,r'_1 \cdot r_2 + a\,r'_2, \ldots)$). Then E is a vector space. Next, define a norm on E as follows: $(r_1, r_2, \ldots = \text{lub}|r_1|$ (noting that the least upper bound on the right exists, since the r_i must be bounded for membership in E). That this is indeed a norm is an easy check. Finally, one must check completeness. To this end, let $x = (r_1, r_2, \ldots)$, $x' = (r'_1, r'_2, \ldots)$, $x'' = \cdots$ be a Cauchy sequence of elements of E. Then the sequence of real numbers, $r_1, r'_1, r''_1, \ldots$ is Cauchy, whence it converges to some real number s_1; the sequence $r_2, r'_2, \ldots$ of real number is Cauchy, whence it converges to some real number s_2: etc. In this way, one constructs a sequence $(s_1, s_2, \ldots)$. Since $(r_1, r_2, \ldots)$ is bounded, and since $x, x', x'', \ldots$ is Cauchy, the sequence $(s_1, s_2, \ldots)$ s bounded: hence, it defines some element y of E. Finally, one observes that, by the construction by which y was obtained, the sequence $x, x', x'', \ldots$ in E converges to y. Thus, we have sketched the proof that E is complete: hence, it is a Banach space.

Example. If, in the example above, one had restricted E to consist only of sequences $(r_1, r_2, \ldots)$ for which $\lim r_i$ (as $i \to \infty$) exists, using the same norm as above, then one would again have obtained a Banach space. Similarly, one could have restricted to $(r_1, r_2, \ldots)$ with $\lim r_1 = 0$, and would still obtain a Banach space. If, however, one had let E consist of sequences $(r_1, r_2, \cdots)$ with the property that, for some N, $r_i = 0$ for $i \geq N$, one would still obtain a vector space E, and one would still have the above norm. However, this would not be a Banach space, for it would not be complete. [For example, let $x = (1, 0, 0, \ldots)$, $x' = (1, 1/2, 0, 0, \ldots)$, $x'' = (1, 1/2, 1/4, 0 \ldots)$, etc, Then $x, x', x'', \cdots$ is a cauchy sequence, but converges to no element of this E (for $(1, 1/2, 1/4, 1/8, \ldots)$ is not in this case an element of E).]

Example. Let E consist of all bounded, real-valued functions on the reals. For f such a function, let $|f| = \text{lub}|f|$. This is a vector space (where addition of functions and multiplication by numbers are point-wise), with norm. In fact, this is a Banach space. [The proof of completeness is identical with that at the bottom of previous page.] If E had consisted only of the continues bounded functions, we would still have a Banach space. Suppose, however, that E had consisted of the differentiable bounded functions. Then, although we would still have a vector space norm, completeness would fail, and we would not obtain a Banach space.

Indeed, consider the sequence of differentiable functions illustrated on the right. These are a Cauchy sequence, but it converges to no differentiable function (the only candidate being the non-differentiable function shown).

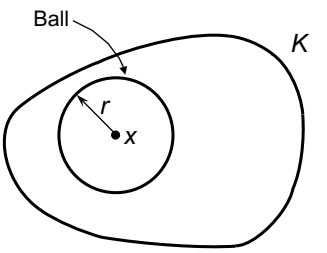


Example Let E consist of all C^n, real-valued functions on the reals, such that the values of the function and of its first n derivatives are all bounded. Let the norm of such a function be $f = \text{lub}|f| + \text{lub}|f'| + \ldots + \text{lub}|f^n|$. This is a Banach space. [If the last term were left off the expression for the norm, completeness would fail.]

Example. Let E consist of all sequences, $(r_1, r_2, \ldots)$, of real numbers, for which the sum $\Sigma|r_i|$ is finite. Let the norm of such a sequence be this sum. This is a Banach space. More generally, fixing any numbers $p \geq 1$, let E consists of $(r_1, r_2, \ldots)$ with $\Sigma|r_i|^p$ finite. Let the norm of an of an element of E be the $p-$th root of this sum. [The taking of root is necessary in order to have $|ax| = |a||x|$.]

We obtain a Banach space.

There is apparently no natural way to make spaces of C^∞ functions into Banach spaces, although there is for analytic functions. It is important that one understand the examples above completely.



We next consider a few additional definitions associated with Banach spaces. Let E, $|\ |$ be a Banach space. For x any vector in E, and r any positive number, the *ball* with *center* x and *radius* r is the subset of E consisting of all vectors y with $|x - y| < r$. The intuitive meaning is that suggested by these terms. [For example, in the second example on 2, the ball with center 0 and radius 1 is essentially the set B of that example.] A subset K of our Banach space is said to be *open* if, given any vector x in K, there is some ball with center x which lies entirely within K. For example, every ball is itself open (a fact whose proof requires use of the second property of a norm). [Those familiar with topology will notice that these are the open sets for a topology on E.] A subset K of E is said to be *closed* if its component in E is open, i.e., if, given any vector x not in K, there is some ball centered at x which does not intersect K.

Now let E be a vector space, and let $|\ |$ and $\{\ \}$ be two norms on E, such that both E, $|\ |$ and E, $\{\ \}$ are Banach spaces. These two norms will be said to be *equivalent* if there are positive numbers a and b such that, for every vector x, $|x| \leq a\{x\}$ and $\{x\} \leq b|x|$. [In intuitive terms, if each norm, after a suitable rescaling, bounds the other.] Note that equivalence of norms is an equivalence relation, and that two norms are equivalent if and only if they define precisely the same open sets on E. It normally happens in practice that the actual numerical values associated with a norm are not relevant to one's particular problem, but rather only that norm up to equivalence. Indeed, it is possible to treat our subject referring only to the open sets (i.e., to equivalence classes of norms), rather than to any particular norm. The resulting treatment is more elegant, and occasionally more awkward. We shall not proceed in this way, however, because it would require a brief retour into topology. As an example of equivalence, we may note the following fact: any two $B's$ in the second example on page 2 give rise to equivalent norms.

Let E, $|\ |$ be a Banach space. A subset F of E is called a *subspace* if the following two conditions are satisfied: i) F is a vector subspace of vector space E (i.e., sums and numerical multiplies of vectors in F are again in F), and ii) F is a closed subset of E. As motivation for this definition, we make the following observation. Let F be a subspace of Banach space E. Then, by condition i), F is itself a vector space. Since every vector in F is also a vector in E, the norm on E also defines a norm on F. We now claim that condition ii) implies that this vector space F, with this norm, is complete (i.e., is a Banach space). [Sketch of proof: Since the norm in F comes from that in E, every Cauchy sequence in F is also a Cauchy sequence in E. Since E is complete, every such sequence converges to some vector in E. Since F is a closed subset of E, this vector must in fact be in F. Hence, F is complete.]

We see here the first of several differences we shall encounter between finite-dimensional and infinite-dimensional Banach spaces. In the finite-dimensional case, condition ii) for a subspace follows already from condition i), i.e., every vector subspace of finite-dimensional E is necessary closed. [Geometrically, e.g., a plane through the origin in Euclidean 3−space is necessarily closed.] This property does not hold, however, for infinite-dimensional Banach spaces:

Example. Let E be the Banach space of the example on page 2 (bounded sequences of real). Let F be the subset of E consisting of all sequences, $(r_1, r_2, \ldots)$ such that $r = 0$ for all $i \geq N$ for some N. [That is, F consists of sequences which are "all zeroes after a while".] Then, since the sum of two sequences in F is again in F, and since a numerical multiple of a sequence in F is again in F, F is a vector subspace of vector space E. We claim, however, that this subset F is not closed. Consider the element

$x = (1, 1/2, 1/4, \ldots)$ of E. This x is not an element of F. If F is to be closed, then there must exist a ball, centered at x, which fails to intersect F. We show that there is no such ball, whence F cannot be closed. Let r be any positive number, and let B be the ball centered at x, with radius r. We must find a vector common to B and F. Set $y = (1, 1/2, 1/4, \ldots, 1/2^n, 0, 0, \ldots)$, where n is chosen sufficiently large that $1/2^n \leq r$. Then $|x - y| = (0, 0, \ldots, 0, 1/2^{n+1}, 1/2^{n+2}, \ldots)$, whence $|x - y| = 1/2^{n+1}$. Thus, the vector y is in the ball B. But y is also in F (since F consists of sequences which are 0 after a while). Hence, this y is common to B and F. [Compare this argument with the argument, in the first part of page 3 that F itself is not a Banach space.]

3. Banach Spaces from Banach Spaces

In this section, we discuss two methods for constructing Banach spaces from Banach spaces. In particular, applied to our examples these methods yield numerous new examples.

Let E and F be Banach spaces. Denote by G the collection of all ordered pairs, (x, y), with x a vector in E and y in F. The set G is made into a vector space by defining addition and scalar multiplication component-wise, i.e., $(x, y) + a(x', y') = (x + ax', y + ay')$. We now wish to make this vector space G into a Banach space, i.e., we wish to define a norm on G. Set $|(x, y)| = |x| + |y|$. The three conditions for a norm are immediate from those conditions in E and F: e.g., $|a(x, y)| = |(ax, ay)| = |ax|+|ay| = |a|\,|x|+|a|\,|y| = |a|(|x|+|y|) = |a||(x, y)|$. Thus, to show that G is a Banach space, we must only check completeness. Let $(x_1, y_1), (x_2, y_2), \ldots$ be a Cauchy sequence in G. Then, for every positive ϵ there is an N such that $|(x_i, y_i)-(x_j, y_j)| \leq \epsilon$ whenever $i, j \geq N$. But $|(x_i, y_i)-(x_j, y_j)| = |x_i-x_j|+|y_i-y_j|$. Hence, $x_1, x_2, \ldots$ is a Cauchy sequence in E (whence it converges to some vector x in E), and $y_1, y_2, \ldots$ is a Cauchy sequence in F (whence it converges to some vector y in F). Consider now the vector (x, y) in G. We have $|(x, y) - (x_i, y_i)| = |x - x_i| + |y - y_i|$. Since $\lim_{i\to\infty}|x-x_i| = \lim_{i\to\infty}|y-y_i| = 0$ (convergence in E and F), we have $\lim_{i\to\infty}|(x, y)-(x_i, y_i)| = 0$. Hence, our sequence converges to (x, y). We have shown that G is complete; hence, is a Banach space. This G is called the *product* of Banach spaces E and F, written $G = E\times F$, Note, for example, that each of E and F may be regarded as a subspace of $E \times F$.

The second construction is like an "inverse" of the first. Let E be a Banach space, and let F be a subspace of E. Then in particular F is a vector subspace of vector space E, and so we may take the quotient of vector spaces; denote the resulting vector space by G. [In more detail: take two vectors x and x' of E as equivalent if their difference, $x - x'$, is an element of subspace F. This is an equivalence relation, and G is the set of equivalence classes. To add two elements of G (equivalence classes), find the equivalence class which includes the sum of one representative from each summand (noting that this is independent of the choice of representative), and similarly for scalar multiplication.] We next introduce a norm on this vector space G. For α any element of G (equivalence class), let $|\alpha|$ be the greatest lower bound of $|x|$, as x varies over the equivalence class α. This is indeed a norm on G. [The third condition for a norm is immediate, and the second is not hard. For the first, first note that always $|\alpha| \geq 0$. If $|\alpha| = 0$, then there is a sequence $x_1, x_2, \ldots$ in α with limit $|x_i| = 0$. Since F is closed, therefore, the vector 0 in E must be in α. Hence, since α has representative 0, we have $\alpha = 0$.] Finally, we show that this G is complete. Let $\alpha_1, \alpha_2, \ldots$ be a Cauchy sequence in G: By choosing a subsequence if necessary, we may suppose that $|\alpha_{i+1} - \alpha_i| \leq 1/2^i$. Choose any vector x_1 in α_1. Choose vector x_2 in $\alpha_2 - \alpha_1$ with $|x_2| \leq |\alpha_2 - \alpha_1| + 1/2$, vector x_3 in $\alpha_3 - \alpha_2$ with $|x_3| \leq |\alpha_3 - \alpha_2| + 1/4$, etc. [These choices are possible by definition of the norm in G.] Thus, for $i \geq 2$ we have $|x_i| \leq 1/2^{i+2}$. Next, set, for each i, $y_i = x_i + x_2 + \ldots + x_i$. Then, for each i, y_i is in α_i. Further, by the bounds above on the norms of the x's, the y_i form a Cauchy sequence in E. Hence, this sequence converges to some vector y in E. Denote by α the equivalence class containing y. Then $|\alpha - \alpha_i| \leq |y - y_i|$ (by definition of the norm in G), and so, since the y_i converge to y in E, the α_i must converge to α in G. We have shown that every Cauchy sequence in G converges to some vector in G, i.e., that G is complete. This Banach space G is called the *quotient* of Banach space E by subspace F, and is written E/F.

The sense in which these two constructions invert each other is this. Let E and F be Banach spaces. Then E is a subspace of the Banach space $E \times F$ (namely, the subspace

consisting of pairs (x, y) with $y = 0$). Hence, we may take the quotient of $E \times F$ by this subspace E. The result is what one expects; $(E \times F)/E = F$

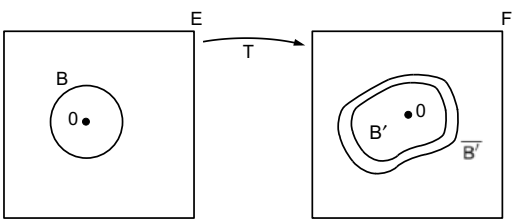
4. Open Mapping Theorem

In this section we establish an important property of Banach spaces. This property is not only useful in applications, but also gives insight into the structure of the definition of a Banach space.

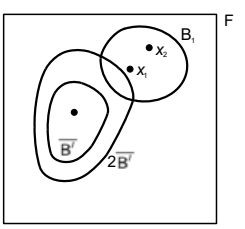
We shall regard the open mapping theorem as a criterion for equivalence of norms.

Theorem (open mapping). Let $E, |\ |$ and $E, \{\ \}$ be two Banach spaces (based on the same vector space E), and suppose that there is a positive number a such that $\{x\} \leq a|x|$ for every vector x. Then these two norms are equivalent, i.e., also $|x| \leq b\{x\}$ for some b.

Proof: First note that no generality is lost by setting $a = 1$. Next, introduce a copy, $F, |\ |$, of Banach space E, $\{\ \}$, and let T denote that mapping from E to F which arises from the identity mapping on E. Then T is linear, one-to-one, onto, and norm-decreasing (i.e., $|T(x)| \leq |x|$ for every x in E). Denote by B the ball with radius one and center 0 in E, by B' the subset $T(B)$ of F, and by $\overline{B'}$ the closure of B' in F (i.e., the set of all vectors in F every sphere centered at which intersects B'). We must show that the subset B' of F contains some ball centered at 0 (for, if $1/b$ is the radius of such a ball, then we shall have $|x| \leq 1$ whenever $|T(x)| \leq 1/b$, which implies immediately that $|T(x)| \leq b|x|$ for every x in E). We divide the proof into three steps.

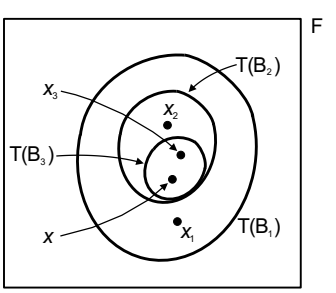


The set $\overline{B'}$ in F contain some ball. Suppose not. We obtain a contradiction. Choose vector x_1 not in $\overline{B'}$. Then, since $\overline{B'}$ is closed, there is some ball B_1, centered at x_1, whose closure $\overline{B_1}$ does not intersect $\overline{B'}$. The set $2\overline{B'}$ (the set of all vectors of the form $2x$ with x in $\overline{B'}$) cannot contain a ball (since $\overline{B'}$ does not), and hence in particular cannot contain B_1. Choose, therefore, vector x_2 in B_1 but not in $2\overline{B'}$. Since $2\overline{B'}$ is closed, there is some ball B_2 centered at x_2 such that $\overline{B_2}$ does not intersect $2\overline{B'}$. Choose B_2 a subset of B_1, and with radius less than half that of B_1. Now $3\overline{B'}$ cannot contain B_2: choose x_3 in B_2 but not in $3\overline{B'}$. Then, choose ball B_3 centered at x_3, not intersecting $3\overline{B'}$, a subset of B_2, and having radius less than half of that of B_2. Continuing in this way, we obtain a sequence $x_1, x_2, \ldots$ in F, which, by construction, must be a Cauchy sequence. Hence, this sequence converges to some vector x in F. By construction, this x cannot be in $\overline{B'}$ (since $\overline{B_1} \cap \overline{B'} = \emptyset$) or in $2B'$ (since $\overline{B_2} \cap 2\overline{B'} = \emptyset$), etc. But this is a contradiction, for $\overline{B'} \cup 2\overline{B'} \cup 3\overline{B'} \ldots = F$ (which follows from the fact that $B \cup 2B \cup 3B \cup \ldots = E$, and that T is onto).



The set $\overline{B'}$ in F contains some ball centered at 0. Let $\overline{B'}$ contain a ball centered at $T(x)$, where x is some vector in E. Then $\overline{B'} - T(x)$ contains a ball centered at 0. But $\overline{B'} - T(x) = \overline{T(B)} - T(x) = \overline{(T(B) - T(x))} = \overline{T(B - x)}$. Choose positive number n such that $B - x$ is a subset of nB. Then $\overline{T(nB)} = n\overline{T(B)} = n\overline{B'}$ contains a ball centered at 0 (since $\overline{T(B - x)}$ does). Hence, $\overline{B'}$ also contains a ball centered at 0.

The set B' in F contains some ball centered at 0. Suppose not: We obtain a contradiction. Let $\overline{B'}$ contain the ball of radius ϵ centered at 0. Since B' contains no ball centered at 0, neither does $3B'$. But $\overline{B'}$ does contain a ball centered at 0: Hence, we may choose a vector $T(x)$ in $\overline{B'}$, but not in $3B'$. Since $T(x)$ is in $\overline{B'}$, we may choose a vector $T(x_1)$ in B' with $|T(x) - T(x_1)| \leq \epsilon/4$.



Denote by B_1 the ball of radius $1/2$ about x_1. Then (since $|T(x) - T(x_1)| \leq \epsilon/4$ and $\overline{T(B_1)}$ contains the ball of radius $\epsilon/2$ about $T(x_1)$) $\overline{T(B_1)}$ contains the ball of radius $\epsilon/4$ about $T(x)$. Hence, we may choose vector $T(x_2)$ in $T(B_1)$ with $|T(x)-T(x_2)| \leq \epsilon/8$. Denote by B_2 the ball of radius $1/4$ about x_2. Then $\overline{T(B_2)}$ contains a ball of radius $\epsilon/8$ about x. Hence, we may choose vector $T(x_3)$ in $T(B_2)$ with $|T(x)-T(x_3)| \leq \epsilon/16$. Denote by B_3 the ball of radius $1/8$ about x_3. Continuing in this way, we obtain a sequence of vectors $x_1, x_2, \ldots$ in E. Since $|T(x) - T(x_i)| \leq \epsilon/2^{i+1}$, the $T(x_i)$ converge to $T(x)$. Since $T(x_i)$ is in $T(B_{i-1})$, x_i must be in B_{i-1}, i.e., we have $|x_i - x_{i-1}| \leq 1/2^{i-1}$. Hence, $x_1, x_2, \ldots$ is a Cauchy sequence in E. Denote by y the vector in E to which it converges. Then, since x_1 is in B (i.e., since $|x_1| < 1$), and since $|x_i - x_{i-1}| \leq 1/2^{i-1}$, we must have $|y| < 3$, i.e., y must be in $3B$. Furthermore, since T is norm-decreasing, and since limit $|y - x_i| = 0$, we must have limit $|T(y) - T(x_i)| = 0$, i.e., the sequence $T(x_i)$ in F converges to $T(y)$. But we have already seen that the $T(x_i)$ converges to $T(x)$: Hence, $T(x) = T(y)$. Thus, since $T(y)$ is in $3B'$, $T(x)$ must also be in $3B'$. This is a contradiction with our original choice of $T(x)$.

This completes the proof of the theorem.

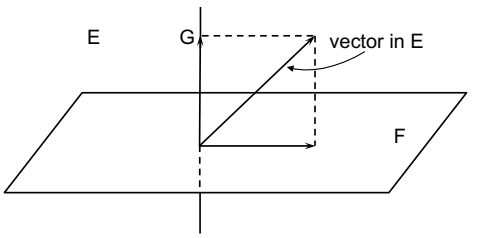
The result above is beautiful and intricate. To see what it means, consider first the finite-dimensional case. Imagine that we have one norm on our vector space, and wish to construct a smaller one. One could imagine obtaining a new norm by "scaling down the original norm by a positive factor, one for each direction (i.e., for each dimension) in the vector space". If there are only a finite number of dimensions, then these "scaling factors" will have a minimum, and so the new norm will also bound (up to a factor) the old one. In the infinite-dimensional case, however, things are different. Now, one could imagine "choosing different factors for the various directions, such that these factors approach zero", thus obtaining a new norm which does not bound the old norm. Suppose, however, that one wants ones new norm to yield a Banach space: Then one must be careful about completeness, for scaling down the old norm "too much" may lead to new Cauchy sequences with nothing to converge to. Is it necessary to actually have the "scaling factors" bounded away from zero in order to avoid destroying completeness? The theorem says yes.

The open mapping theorem is somewhat analogous to a theorem in topology: Given a compact, Hausdorff topological space, then no finer and no coarser topology can be both compact and Hausdorff. The open mapping theorem says, similarly: Given a Banach space, then no larger and no smaller norm (except for equivalence) can also give a Banach space. In fact, whenever, in an argument about Banach spaces, one is tempted to try to use compactness (which is almost never available in this subject), one should try instead to apply the open mapping theorem.

5. Splitting

An important issue about a subspace of a Banach space is that of whether or not it has the property called splitting. In this section, we introduce this property, show that it is always satisfied in certain cases, and give an example in which it is not satisfied.

Let E be a Banach space, and F a subspace of E. A subspace G of E is said to be *complementary* to F if the following property is satisfied: every vector in E can be written in one and only one way as the sum of one vector in F and one in G.



Example. Let E be the Banach space of all sequences, $(r_1, r_2, \ldots)$, of real numbers, such that the r_i converge, and let the norm of such a sequence be the least upper bound of the absolute values of its entries. Let E be the subspace consisting of sequences which converge to zero, and let G be the subspace consisting of constant sequences. Then F and G are complementary.

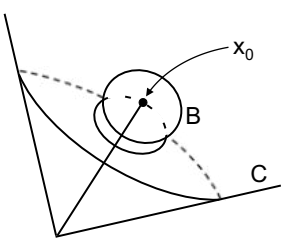
Let F and G be complementary subspaces of Banach space E. We define a mapping from G to the quotient (Banach) space E/F as follows: This mapping takes the vector x in G to the equivalence class (element of E/F) which contains x. This mapping is clearly linear, one-to-one, and onto. Furthermore, this mapping is norm-decreasing (since the norm of x in G is just the norm of x in E, while the norm of an equivalence class (in E/F) is the greatest lower bound of the norms of the representatives (including x) of that class). By the open mapping theorem, therefore, this mapping is an isomorphism of Banach spaces. Thus, if G is complementary to F in E, then G represents "a realization of E/F in E", The Banach space E/F starts out as just an abstract Banach space: it does not "live" in E. The finding of a complementary subspace (to F) "realizes E/F".

Let E be a Banach space, and F a subspace of E. The subspace F is said to *split* if there exists in E a subspace complementary to F. The purpose of this section is to understand this definition.

We first show that "very small" subspaces do split.

Theorem. Let F be a one-dimensional subspace of Banach space F. Then

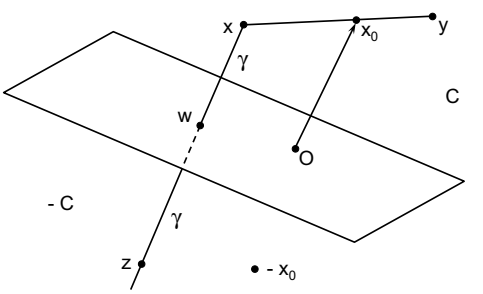
Proof: Choose vector x_0, with $|x_0| = 1$, in F. Denote by B the ball with center x_0 and radius $1/2$. Next, denote by ζ the collection of all subsets C of E having the following properties: i) C is convex (i.e., the line segment joining any two vectors in C lies within C), ii) C is radial (i.e., if x is in C and a is a positive number, then ax is in C, iii) C contains the ball B, and iv) C does not contain the vector $-x_0$. There certainly exists at least one subset in ζ, namely, that consisting of all positive multiples of vectors in B. Partially order the set ζ by inclusion (i.e., $C_1 \leq C_2$ if $C_1 \subset C_2$). We next note that this partially ordered set ζ satisfies the condition for Zorn's Lemma (namely, that every totally ordered subset of ζ is bounded above), for, given a totally ordered subset of ζ, their union gives such a bound. [Conditions i)–iv) for this union are all immediate from those conditions for the C's in the totally ordered subset.] So, by Zorn's Lemma, there exists a maximal element, C, of ζ.



We next obtain two properties of this set C. First, $C \cup (-C) = E$. Indeed, if there were some vector x in E, with x in neither C nor $-C$, then we could consider the set C', consisting of the union of all line segments joining a positive multiple of x to a vector in

C. This C' would clearly satisfy conditions i), ii), and iii) above. It would, furthermore, satisfy condition iv), for if $-x_0$ were in C', say $-x_0 = ax + y$ (a positive, y in C), then we would have $x = -1/a(x_0 + y)$, whence x would be in $-C$. But the existence of this C' (an element of ζ, since it satisfies all four conditions for membership) would violate maximality of C. Hence $C \cup (-C) = E$. The second property is that C is closed. Indeed, if not, we could let C' be the closure of C. Then this C' clearly satisfies conditions i), ii), and iii) above. Furthermore, it would satisfy condition iv), for, were $-x_0$ in C', there would have to be some vector y in C with $|y + x_0| \leq 1/4$. But $-y - 1/8\, x_0$ must be in C (for $|(-y - 1/8\, x - 0) - x_0| = |y + x_0 + 1/8\, x_0| \leq |y + x_0| + |1/8\, x_0| \leq 3/8$, whence $-y - 1/8\, x_0$ is already in B). Hence, since C is convex, $1/2(y) + 1/2(-y - 1/8\, x_0) = -1/16\, x_0$ would be in C, whence $-x_0$ would be in C; a contradiction. Thus, the closure C' of C must also be an element of ζ. This violates maximality of C unless C is already closed (so $C' = C$).

Now set $G = C \cap (-C)$. This subset G of E is a vector subspace of vector space E (since, for x and y in G, $ax + (z - a)y$ is in G (for $0 \leq a \leq 1$; convexity of C and $-C$), ax is in G (for $a > 0$; radialness of C and $-C$), and $-x$ is in G (since, for x in G, x is in both C and $-C$, whence $-x$ is also in both). Furthermore, this subset G is closed (as the intersection of closed subsets C and $-C$). Thus, G is actually a subspace of Banach space E. This subspace G is our candidate for a subspace complementary to F.



First note that no vector can be written in more than one way as a linear combination of x_0 and a vector in G, for, were this possible, x_0 itself would have to be in G (a contradiction, since x_0 is not in $-C$). Thus, we have only to show that every vector in E can be written in some way as a linear combination of x_0 and a vector in G. Let x be any vector in E (say, x in C). Denote by γ the straight line in E through x parallel to the vector x_0 (i.e., the set of all vectors of the form x plus a multiple of x_0). Choose vector y a little beyond x_0 on the line joining x and x_0. By making the "little" small enough, we can have y in B (i.e., we can have $|y - x_0| \leq 1/2$). Clearly, there is a point z on the line γ such that the line segment joining z and y contains $-x_0$. Then this z cannot be in C (for y, being in B, is necessary in C, and if z were also in C then, by convexity, $-x_0$ would be in C, contradicting definition of C). Hence, the line γ has some points in C and some in $-C$. Therefore, there must exist some point w on γ "at the interface" (i.e., such that, given any ϵ, the intersection of γ with the ball of radius ϵ about w intersects both C and $-C$). Since C and $-C$ are closed, this w must be in both C and $-C$, i.e., w must be in G. But x can be written as a linear combination of x_0 and w. Thus, the subspaces G and F are indeed complementary, completing the proof.

It is immediate from this result that every finite-dimensional subspace of a Banach space split. Indeed, let F be such a subspace, and let $x_1, \ldots, x_n$ be a basis for F. Denote by $G_1, \ldots, G_n$ a subspace complementary to that spanned by $x_1, \ldots, x_n$, respectively. Then $G \cap \ldots \cap G_n$ is a subspace complementary to F. It is also easy to show that any subspace of finite co-dimension (codimension of F in E equals dimension of R/F, i.e., is the dimension of "what is left over in E after F") splits. Indeed, choose basis $\alpha_1, \ldots \alpha_n$ for E/F, and representatives $x_1, \ldots, x_n$ or these equivalence classes in E. Let G be the subspace spanned by $x_1, \ldots, x_n$. Then, G is complementary to F. [This same argument does not work when E/F is infinite-dimensional, for then the G, so constructed, may not be closed.]

We thus conclude that, given Banach space E, "very large" subspaces (finite co-dimension) and "very small" subspaces (finite dimension) split. Thus, if any subspaces are to fail to split, they must be of "intermediate size" (infinite in both dimension and co-dimension). It is perhaps not even obvious that there exists an example of a subspace which fails to split. It turns out that such examples do exist, but that it is surprisingly dif-

ficult to display one. We now give such an example (an example we shall not use again, but which is nonetheless perhaps worthwhile, since it gives some insight into the nature of the notion of splitting).

Examples. For purposes of this example, we define a functional on Banach space E as a linear mapping f from E to the reals such that $|f(x)| \leq |x|$ for every x in E.

Denote by E^1 the Banach space consisting of all sequences, $(r_1, r_2, \ldots)$, of real numbers the sum of whose absolute values, $\Sigma|r_i|$ is finite, where the norm of such a sequence is this sum. We next wish to claim that this Banach space E^1 has the following property: Given any sequence, $x_1, x_2, \ldots$, of unit vectors in E^1, there is some functional f on E^1 such that the sequence of numbers $f(x_1), f(x_2). \ldots$ fails to converge to zero. The proof is as follows. For any finite subset L of the positive integers, and any positive integer i, denote by S^i_L the real number obtained by taking the sum $\Sigma_L r_j$, where $x_i = (r_1, r_2, \ldots)$. If, for any finite L, we did not have $\lim_{i\to\infty} S^i_L = 0$, then we would be done: let f be the functional which assigns to $x = (s_1.s_2, \ldots)$ the number $f(x) = \Sigma_L s_j$. So, we may suppose that $\lim_{i\to\infty} S^i_L = 0$ for every L. Choose finite L^1 such that $|S^1_{L^1}| \geq 1/4$. Then, find integer n_1 such that $|S^j_{L^1}| \leq 1/16$ whenever $j \geq n_1$. Next, choose finite L^2, disjoint from L^1, such that $|S^{n_1}_{L^2}| \geq 1/4$.

Then, find integer $n_2 > n_1$ such that $|S^j_{L^2}| \leq 1/32$ whenever $j \geq n_2$. Next, choose finite L^3, disjoint from L^1, such that $|S^{n_2}_{L^3}| \geq 1/4$. Continuing in this way, we obtain disjoint finite sets $L^1, L^2, \ldots$. We next introduce a sequence $(a_1, a_2, \ldots)$ of real numbers according o the following rule: $a_i = 0$ if i is none of $L^1, L^2, \ldots$ $a_i = +1$ if i is in some L^j with $S^{n_{j-1}}_{L^j}$ positive, and $a_i = -1$ if i is in some L^j with $S^{n_{j-1}}_{L^j}$ negative. Now let f be the following functional on Banach space E^1, for $x = (r_1, r_2, \ldots)$ in E^1, $f(x) = r_1a_1 + r_2a_2, + \ldots$ (noting that this sum converges, since the sum of the absolute values of the r_i is finite, and the $|a_i|$ are bounded). But, by construction, $|f(x_i)| \geq 1/8$ for i one of the n_j. Hence, $f(x_1), f(x_2), \ldots$ cannot converge to zero.

We next denote by E^2 the Banach space consisting of all sequences $(r_1, r_2, \ldots)$ of real numbers for which the sum $\Sigma|r_i|^2$ is finite, where the norm of such a sequence is the square root of this sum. We claim that the Banach space E^2 fails to satisfy the property that we just showed for E^1. That is: there exists a sequence $x_1, x_2, \ldots$ of unit vectors in E^2 such that $f(x_i)$ goes to zero for every functional f. Indeed, let $x_1 = (1, 0, 0, \ldots)$, $x_2 = (0, 1, 0, 0, \ldots)$, $x_3 = (0, 0, 1, 0, \ldots)$, etc. Set $a_i = f(x_i)$, where f is some functional. then for any numbers $r_1, r_2, \ldots$ with $\Sigma|r_i|^2$ finite, the sequence in E whose n^{th} term is $y_n = r_1x_1 + \ldots + r_nx_n$ is Cauchy, whence it converges to some vector y in E. Then, since f is a functional, the sequence $f(y_i)$ of real numbers must converge, whence the sum $\Sigma r_i a_i$ must converge. That is to say, the numbers a_i must be such that $\Sigma r_i a_i$ is finite whenever $\Sigma|r_i|^2$ is finite. But this is possible only if $\Sigma|a_i|^2$ is finite. But this, in turn, requires that the a_i converge to zero. We conclude: $f(x_i)$ approaches zero. But this is what we wanted to show.

Thus, we now have two Banach spaces which differ in a certain respect. [In particular, we know that there exists no isomorphism from E^1 to E^2.]

The next step is the introduction of a certain mapping from E^1 to E^2. Let $x_1, x_2, \ldots$ be the sequence (ordered in some way) of all vectors in E^2 of the following form: $a(r_1, r_2, \ldots, r_n, 0, 0, \ldots)$, where $r_1, \ldots, r_n$ are rational, and the number a is so chosen that these are unit vectors in E^2. The mapping ψ is now defined as follows. Given any element $x = (s_1, s_2, \ldots)$ in E^1, let $\psi(x)$ be that element of E^2 which is the limit y of the Cauchy sequence in E^2 whose n^{th} term is $y_n = s_1x_1 + \ldots + s_nx_n$ (noting that $y_1, y_2, \ldots$ is indeed a Cauchy sequence since $|x_i| = 1$ for every i, while $\Sigma|s_j|$ is finite). This mapping ψ is clearly linear. Furthermore, it is onto E^2 (for finite linear combinations of the x_i are dense in E^2). Furthermore, this mapping is even norm-decreasing (for, since the x_i are unit, $|s_1x_1 + \ldots + s_nx_n| \leq |s-1| + \ldots + |s_n|$).

Denote by K the kernel of the mapping ψ, i.e., the set of all vectors x in E^1 with $\psi(x) = 0$. Then, since ψ is linear, K is a vector subspace of vector space E^1. Further,

since ψ is norm-decreasing, K is a closed subset of E^1 (for, given sequence $z_1, z_2, \ldots$ in K, converging to z, then $|\psi(z) - \psi(z_i)| \leq |z - z_i|$, whence, since the z_i converge to z, the $\psi(z_i)$ converge to $\psi(z)$. But $\psi(z_i) = 0$ and so we must have $\psi(z) = 0$, i.e., x must also be in K.)

Thus, we now have Banach space E^1, and subspace K. We claim, finally, that this subspace of E^1 does not split. Suppose that it did, say with complementary subspace G: We obtain a contradiction. Since the mapping ψ is onto, the subspace G must be isomorphic with Banach space E^2. Choose a sequence of vectors in E^2, each unit, but such that for any functional f on E^2, f, applied to this sequence, approaches zero. Since E^2 is isomorphic with G, we obtain a similar sequence in E^1. The resulting sequence in $E^1, w_1, w_2, \ldots$ will have the property that, for every functional f on E^1, limit $f(w_i) = 0$ (since every functional on E^1 can also be regarded as a functional on the subspace G). Furthermore, the norms of the vectors $w_1, \ldots$ will be bounded below (since the w_i arose originally from a sequence of unit vectors in E^2). But, (after rescaling the w_i) so that they are unit) this contradicts the property of E^1 that we showed at the beginning: whenever $f(w_i) \to 0$ for every f, $|w_i| \to 0$. This contradiction establishes that the subspace K of E^1 has no complementary subspace, completing the example.

The subspace of converging sequences in the Banach space of bounded sequences also has no complementary subspace. Every subspace of E^2 has a complementary subspace.

The general rule seems to be that either one can discover with relative ease some explicit complementary subspace, or none exists. Note, incidentally, that, in the finite-dimensional case, every subspace splits. It is for this reason that splitting is not normally mentioned in the finite-dimensional case.

6. Bounded Linear Mappings

The "structure-preserving" (and, therefore, the useful, and interesting) mappings between Banach spaces are what are called bounded linear mappings. In this section, we introduce these, and give some examples and properties.

Let E and F be Banach spaces. A mapping T from E to F is called a *bounded linear mapping* provided i) T is a linear mapping of vector spaces (i.e., for every x and y in E, and every real number b, $T(x+by) = T(x)+bT(y)$), and ii)T is bounded (i.e., there exists a number a such that, for every x in E, $|T(x)| \leq a|x|$).

We have already used this notion, although not this term, several times. what was called a functional on page 12 was precisely a bounded linear mapping from E to $\mathbb{R}$. Banach spaces $E, |\ |$ and $E, \{\ \}$ are equivalent (page 4) if and only if both the identity mapping from $E, |\ |$ to $E, \{\ \}$ and its inverse are bounded linear mappings. the open mapping theorem can be restated thus: if a bounded linear mapping from one Banach space to another is one-to-one and onto (so its inverse exists), then that inverse is also a bounded linear mapping.

For T a bounded linear mapping from Banach space E to Banach space F, the smallest (necessarily non-negative) number a such that $T|(x)| \leq a|x|$ for every x is called the *norm* or *bound* of T, and is written $|T|$.

Example. Let E be the Banach space of bounded sequences (with norm the least upper bound of absolute values of entries), and F the Banach space of convergent sequences (same norm). The identity mapping from F to E (every element of F is automatically an element of E) is a bounded linear mapping. The mapping from F to $\mathbb{R}$ which assigns to each sequence its limit is a bounded linear mapping. The mapping from E to $\mathbb{R}$ which assigns to $(r_1, r_2, \ldots)$ the number r_{137} is a bounded linear mapping. Each of these mappings has norm one.

Example. Let E be the Banach space of sequences the sum of the absolute values of whose entries converges, and let F be the same, with the sum of the squares of absolute values. The identity mapping from E to F is a bounded linear mapping, with norm one. The mapping ψ in the beginning of page 12 is a bounded linear mapping: the remarks of that paragraph show that its norm less than or equal to one.

Example. Let E be the Banach space of C^n functions on the reals, the value of whose first n derivatives are bounded (page 3). Let F be the Banach space of C^{n-1} functions. Let T be the mapping "take the derivative" from E to F. Then T is a bounded linear mapping (with norm one). The mapping from E to R given by "evaluate the derivative at real number 7" is also a bounded linear mapping with norm one (note that in this case, however, there is no nonzero x with $|T(x)| = |T|\,|x|$).

In the finite-dimensional case, the second condition for a bounded linear mapping is unnecessary: every linear mapping from one finite-dimensional Banach space to another is necessarily bounded. It turns out that boundedness does not follow from linearity in the infinite-dimensional case, but examples seem to be rather difficult to display.

Example. Let each of E and F be the Banach space of bounded sequences of real numbers. Denote by A the subset of E consisting of sequences which are zeros after a while. Then A is a vector subspace of vector space E (but not a Banach subspace). For $(r_1, \ldots, r_n, 0, 0, \ldots)$ in A, set $T(r_1, \ldots, r_n, 0, \ldots = (r_1, 2r_2, 3r_3, \ldots nr_n, 0, \ldots)$. Clearly, if we can extend the action of this T linearly to all of E, we shall have our counterexample. Consider the collection of all vector subspace E having in common with A only the zero vector. Ordering by inclusion, we have that the hypothesis of Zorn's lemma is satisfied. Let B be a maximal element: Then, by maximality, every vector in E can be written in one and only one way as the sum of one vector in A and one in B. For $x = y + z$ in E, with y in A and z in B, set $T(x) = T(y)$. This is the desired extension.

Finally, we consider candidates for subspaces defined by a bounded linear mapping. Let E and F be Banach spaces, and let T be a bounded linear mapping from E to F. We denote by Ker T, the *kernel* of T, the set of all vectors x in E such that $T(x) = 0$. Then Ker T is clearly a vector subspace of vector space E. We claim, furthermore, that Ker T is closed. Indeed, let $x_1, x_2, \ldots$ be a sequence in Ker T, converging to x in E. Then, since $|T(x) - T(x_i)| \leq |T|\,|x - x_i|$, the sequence $T(x_1), T(x_2), \ldots$ in F converges to $T(x)$ in F. But the $T(x_i)$ are zero, whence $T(x) = 0$, whence x is also in Ker T. Thus, Ker T is a subspace of E. Next, denote by Im T, the *image* under T, the set of all vectors in F of the form $T(x)$ for some x in E. Then Im T is clearly a vector subspace of vector space F. However, Im T is not in general closed.

Example. Let E be the Banach space of bounded sequences, and F that of convergent sequences. Let T be the bounded linear mapping from E to F with action $T(r_1, r_2, \ldots) = (r_1, r_2/2, r_3/3, \ldots)$ (so $|T| = 1$). Then the vector $(1/\sqrt{1}, 1/\sqrt{2}, 1/\sqrt{3}, \ldots)$ in F is certainly in the closure of Im T, but is the image of no vector in E (for the only candidate is $\sqrt{1}, \sqrt{2}, \sqrt{3}, \ldots$), which is not in E. In the finite-dimensional case, Im T, as a vector subspace, is always closed.

We shall next be concerned with spaces of bounded linear mappings. Fix Banach spaces E and F. We denote by $\mathcal{L}(E; F)$ the set of all bounded linear mappings from E to F. We define addition and multiplication by reals within this set $\mathcal{L}(E; F)$ thus: For T and T' in $\mathcal{L}(E : F)$, and a a real number, let $T + a\,T'$ be that mapping with $(T + a\,T')(x) = T(x) + a\,T'(x)$. [Note that the linear mapping $T + a\,T'$, so defined, is indeed bounded, for $|(T + a\,T')(x)| = |T(x) + a\,T'(x)| \leq |T(x)| + |a|\ |T'(x)| \leq (|T| + |a|\ |T'|)\ |x|$. Thus, $|T + a\,T'| \leq |T| + |a|\ |T'|$.] With these operations, the set $\mathcal{L}(E; F)$ becomes a vector space. We next note that "take the norm" is a norm on the vector space $\mathcal{L}(E; F)$, the three properties for a norm being immediate from the little formula just derived. In fact, this vector space with norm, $\mathcal{L}(E; F)$ is even a Banach space, i.e., it is complete under its norm. Proof: Let $T_1, T_2, \ldots$ be a Cauchy sequence in $\mathcal{L}(E; F)$. Then for each vector x in E, we have, since $|T_i(x) - T_j(x)| \leq |T_i - T_j|\ |x|$, that $T_1(x), T_2(x), \ldots$ is a Cauchy sequence in F. Hence, this sequence converges to some vector in F, which we denote $T(x)$, thus defining a (clearly linear) mapping T from E to F. We next note that this T is bounded, for, for any x in E, $|T(x)| = \lim|T_i(x)| = \lim|T_1(x) - (T_1 - T_i)(x)| \leq |T_1(x)| + \lim|(T_1 - T_i)(x)| \leq (|T_1| + \lim|T_1 - T_i|)\,|x|$. We show that the T_i converge to T in $\mathcal{L}(E; F)$. Given positive ϵ, choose i so that $|T_i - T_j| \leq \epsilon$ for $j \geq i$. Then $|(T - T_i)(x)| \leq \operatorname{limit}_{j\to\infty}|(T - T_j)(x) + (T_j - T_i)(x)| \leq \lim|(T - T_j)(x)| + \lim|(T_j - T_i)(x)| \leq \epsilon|x|$, whence $|T - T_i| \leq \epsilon$. Thus, we conclude that $\mathcal{L}(E; F)$ is a Banach space. This construction yields many new Banach spaces from those we have considered already.

One generalizes the notion of a bounded linear mapping to several independent variables as follows. Let $E_1, \ldots, E_n$ and F be Banach spaces. A mapping T which assigns to an $n-$tuple of vectors, one from each of the E_i, a vector in F is called a *bounded multilinear mapping* provided, i) T is multilinear, i.e., linear in each variable separately (e.g., $T(x_1, x_2 + ax'_2, x_3, \ldots, x_n) = T(x_1, x_2, \ldots, x_n) + aT(x_1, x'_2, \ldots x_n)$), and ii) T is bounded, i.e., there is a number a such that $|T(x_1, \ldots, x_n)| \leq a\ |x_1||x_2| \ldots |x_n|$. The smallest such number is called the *bound*, or *norm*, of T, and is again denoted $|T|$. The various properties of bounded linear mappings carry over almost immediately to similar properties of bounded multilinear mappings. In particular, the set of all such form $E_1, \ldots, E_n$ to F, denoted $\mathcal{L}(E_1, \ldots, E_n; F)$, is a Banach space.

There is a certain sense, which we now explain, in which multilinear mappings are extraneous. Let E, F, and G be Banach spaces. We define a mapping ψ from the Banach space $\mathcal{L}(E, F; G)$ to $\mathcal{L}(E; \mathcal{L}(F; G))$ as follows: For T in $\mathcal{L}(E, F; G)$, let $\psi(T)$ be that element of $\mathcal{L}(E; \mathcal{L}(F; G))$ which sends the vector x in E to that element of $\mathcal{L}(F; G)$ whose action on vector y in F is the vector $T(x, y)$ in G. This mapping is clearly linear. It is, in fact, invertible: Indeed, ψ^{-1} sends the element W of $\mathcal{L}(E; \mathcal{L}(F; G))$ to that element pf $\mathcal{L}(E, F; G)$ whose action on x, y (x in E,; y in F) is the element $[W(x)](y)$ of G

(nothing that $W(x)$ is an element of $\mathcal{L}(F;G)$). It is furthermore immediate from the definitions (although confusing in detail) that ψ is norm-preserving. Thus, the Banach spaces $\mathcal{L}(E,F;G)$ and $\mathcal{L}(E;\mathcal{L}(F;G))$ are equivalent. Those who like $\mathcal{L}$'s will note that our ψ is a preferred element of $\mathcal{Z}(\mathcal{L}(E,F;G);\mathcal{L}(E;\mathcal{L}(F;G))$. Spaces of multilinear mappings can thus always be expressed in terms of iterated $\mathcal{L}$'s.

We conclude this section with the introduction of two additional "preferred objects". Let E and F be Banach spaces. Denote by α the following element of $\mathcal{L}(E,\mathcal{L}(E;F);F)$: For x in E and T in $\mathcal{L}(E;F)$, $\alpha(x,T) = T(x)$ (an element of F). Thus, α is the multilinear mapping whose action is "action of an element of $\mathcal{L}(E;F)$ on an element of E". Multilinearity of α is obvious; we have only to show roundedness. But we have $|\alpha(x,T)| = |T(x)| \leq |T|\ |x|$. Thus, not only is α bounded, but its norm is one. For the second object, let E, F and G be Banach spaces. Let β denote the following element of $\mathcal{L}(\mathcal{L}(E;F),\mathcal{L}(F;G);\mathcal{L}(E;G))$: For T in $\mathcal{L}(E;F)$ and U in $\mathcal{L}(F;G)$, let $\beta(T,U) = U \cdot T$ (composition of mappings: a mapping from E to G). We must first show that the linear mapping $U \cdot T$ is actually in $\mathcal{L}(E;G)$, i.e., that it is bounded. For x in E, $|(U \cdot T)(x)| = |U|\,[T(x)]|| \leq |U|\ |T(x)| \leq |U|\ |T|\,|x|$, whence $|U| \cdot |T| \leq |U|\ |T|$. Thus, $U \cdot T$ is indeed in $\mathcal{L}(E;G)$, and so our mapping β is indeed well-defined. It is obviously that this β is multilinear. Thus, we have only to show that β is bounded. But for T in $\mathcal{L}(E;F)$ and U in $\mathcal{L}(F;G)$, we have $|\beta(T,U)| = |U \cdot T| = |U \cdot T| \leq |U|\ |T|$. Thus, is indeed bounded, and in fact we have $|\beta| = 1$. This β, of course, just performs for us the operation of "composition of bounded linear mappings". That β is itself a bounded multilinear mapping expresses the basic properties of composition.

What is so nice about all this is that everything in sight is a Banach space: Everything one tries to do with Banach spaces yields just other Banach spaces. The result is that one has to learn in detail but a single kind of mathematical object.

7. Derivatives

Elementary differential calculus (of several variables) is of course carried out in $\mathbb{R}^n$. But $\mathbb{R}^n$ is a particular example of a Banach space. In this section, we shall see that finite-dimensionality plays essentially no role in differential calculus: We shall, indeed, repeat the basic ideas of this subject for Banach spaces.

Fix Banach spaces E and F, Fix an open subset U of E (so, for every point x in U, there is a ball centered at x in U). Let f be a mapping from the set U to the set F. [In the finite-dimensional case, with E $n-$dimensional and F $m-$dimensional, f is represented as m functions of n real variable.] [Note that f is only defined on U. whenever we speak of the action of f on a ball centered at a point of U, we shall suppose implicitly that this ball is small enough so that it is within U.]

Fix a point x_0 of U. The mapping f will be said to be *continuous* at x_0 provided that, for every positive number ϵ, there is a positive number δ such that $|f(x) - f(x_0)| \leq \epsilon$, whenever $|x - x_0| \leq \delta$. This f is said to be just *continuous*, or C^0, if it is continuous at every point of U.

We next wish to get hold of the "rate of change of $f(x)$ with x". We first introduce the notion of "zero rate of change". The mapping f is said to be *tangent* at x_0 provided that, for every positive number ϵ, there is a positive number δ such that $|f(x)-f(x_0)| \leq \epsilon|x-x_0|$ whenever $|x-x_0| \leq \delta$. Thus, f is tangent at x_0 provided "the deviation of $f(x)$ from $f(x_0)$ relative to the deviation of x from x_0 becomes as small as we wish (ϵ) whenever x is sufficiently close (δ) to x_0." Note that, if f is tangent at x_0, then f is automatically continuous at x_0 (although the converse, as we shall see shortly, is false).

The mapping f is said to be *differentiable* at x_0 provided that there exists a bounded linear mapping T from E to F such that $f(x) - f(x_0) - T(x - x_0)$ is tangent at x_0. In other words, we require that, given any positive ϵ, there is a positive δ such that $|f(x) - f(x_0) - T(x-x_0)|/|x-x_0| \leq \epsilon$ whenever $x \neq x_0$ and $|x-x_0| \leq \delta$. Thus, for differentiability we require that "$f - f(x_0)$ can be approximated, up to tangency, by a bounded linear mapping". This T, called the *derivative* of f at x_0, is written $Df(x_0)$, so $Df(x_0)$ is an element of $\mathcal{L}(E;F)$. [This is what we expect. In the finite-dimensional case, $Df(x_0)$ is represented by n partial derivatives of the m functions representing f. But this $n \times m$ matrix can be regarded as a linear mapping from $n-$dimensional E to $m-$dimensional F.] We shall see shortly that the derivative is unique if it exists. If f is differentiable at every point x_0 of U, then f is said to be just *differentiable*. In this case Df assigns, to each point x_0 of U, an element $Df(x_0)$ of $\mathcal{L}(E;F)$. Thus, Df is a mapping from U (an open subset of Banach space E) to $\mathcal{L}(E;F)$ (a Banach space). If this mapping Df is continuous (examples in finite dimensions show that it need not always be), then f is said to be *continuously differentiable*, or C^1.

In the finite-dimensional case, continuous, differentiable, and continuously differentiable reduce, of course, to their usual meanings.

Example. Let E and F be Banach spaces, and fix bounded linear mapping T from E to F. Set $U = E$, and let f be the mapping from U to F with $f(x) = T(x)$. Then, for any given x_0 in U, we have for all x $|f(x) - f(x_0)| \leq |T|\,|x - x_0|$. Thus, f is continuous (given ϵ, choose $\delta = \epsilon/|T|$). By definition of the norm, we have that, for fixed x_0 in U and for any number $a > 1$, there is an x in U such that $|f(x) - f(x_0)| \geq a|T|\,|x - x_0|$. But by linearity this inequality continues to hold if x is replaced by any vector on the line joining x and x_0. Hence, f cannot be tangent at x_0 unless f is the zero linear mapping (a mapping which is necessarily tangent at every x_0). We next verify that f is differentiable at x_0. Indeed, $f(x)-f(x_0)-T(x-x_0) = 0$ for all x, whence the left side is certainly tangent at x_0. Hence, $Df(x_0) = T$. Thus, Df assigns to the element x_0 of U the fixed element T of $\mathcal{L}(E;F)$. Since this (constant) mapping Df from U to $\mathcal{L}(E;F)$ is certainly continuous, our f is C^1.

Example. Let f and f' both be continuous. Then so is $f + f'$ (with action $(f + f')(x) = f(x)+f'(x)$). Indeed, given ϵ, find δ_1 such that $|f(x)-f(x_0)| \leq \epsilon/2$ whenever $|x-x_0| \leq \delta_1$, and δ_2 such that $|f'(x) - f'(x_0)| \leq \epsilon/2$ whenever $|x - x_0| \leq \delta_2$. Then, whenever $|x - x_0| \leq \min(\delta_1, \delta_2)$, we have $|(f + f')(x) - (f + f')(x_0)| \leq \epsilon$. By a similar argument, the sum of two functions tangent at x_0 is again tangent at x_0. Next, suppose that each of f and f' is differentiable at x_0. Then, since the functions with action $f(x) - f(x_0) - Df(x_)(x - x_0)$ and $f'(x) - f'(x_0) - Df'(x_0)(x - x_0)$ are both tangent at x_0, so is their sum, $(f + f')(x) - (f + f')(x_0) - (Df(x_0) + Df'(x_0))(x - x_0)$. Hence, $f + f'$ is also differentiable at x_0, and, furthermore, $D(f + f')(x_0) = Df(x_0) + Df'(x_0)$. But now, since the sum of continuous functions is continuous, it is immediate from $D(f + f') = Df + Df'$ that the sum of two C^1 functions is C^1.

It follows from these two example that the derivative $Df(x_0)$, if it exists, is unique. Indeed, let T and T' be two (so each is in $\mathcal{L}(E; F)$). Then both $f(x)-f(x)-T(x-x_0)$ and $f(x) - f(x_0) - T'(x - x_0)$ are tangent at x_0, whence their difference, the mapping $T - T'$, is also tangent. But the only bounded linear mapping which is tangent at x_0 is the zero one, whence $T = T'$.

Example. Let each of E and F be the Banach space of continuous, bounded, real-valued functions on the reals. Let $U = E$, and let f be the mapping from U to F which function φ to φ^2 (noting that the square of a continuous bounded function is another). Fix element φ_0 of U. Then f continuous at φ_0. Indeed, $|f(\varphi)-f(\varphi_0)| = |\varphi^2-\varphi_0^2| = |(\varphi-\varphi_0+2\varphi_0)(\varphi-\varphi_0)| \leq (|\varphi - \varphi_0| + 2|\varphi_0|)|\varphi - \varphi_0|$. Hence, by making $|\varphi - \varphi_0|$ sufficiently small we also make $|f(\varphi) - f(\varphi_0)|$ as small as we wish. We next show that this f is in fact differentiable at φ_0. As a general rule, it is difficult ro show differentiability without first making a guess as to what the derivative is to be. Taking our cue from elementary calculus, we guess as follows: Let α be the bounded linear mapping from E to F which sends element φ of E to element $\alpha(\varphi) = 2\varphi\varphi_0$ of F. Then we have $f(\varphi)-f(\varphi_0)-\alpha(\varphi-\varphi_0) = \varphi^2-\varphi_0{}^2-2\varphi_0(\varphi-\varphi_0) = (\varphi - \varphi_0)^2$. Thus, $|f(\varphi) - f(\varphi_0) - \alpha(\varphi - \varphi_0)| \leq |\varphi - \varphi_0|^2$. Clearly, then, (choose $\delta = \epsilon$), the mapping on the left is tangent at φ_0. We conclude that f is differentiable, and $Df(\varphi_0) = \alpha$. Finally, we claim that this mapping Df from U to $\mathcal{L}(E; F)$ is itself continuous. Indeed, $|Df(\varphi) - Df(\varphi')| = |Df(\varphi - \varphi')| \leq 2|\varphi - \varphi'|$. Thus, the mapping f is C^1.

Example. Let E be any Banach space, and let F be the Banach space of reals. Set $U = E$, and let f be the mapping from U to F which sends x in U to $f(x) = |x|$. Then, given x_0 in U, f is continuous at x_0, for $|f(x)-f(x_0)| = ||x|-|x_0|| \leq |x-x_0|$ [Choose $\delta = \epsilon$.]. However, f need not necessarily be differentiable at x_0, even for $x_0 \neq 0$. For example, let E be $\mathbb{R}^2$, with norm $|(r, r')| = \max(|r|, |r'|)$. Then f is not differentiable, e.g., at $x_0 = (1, 1)$.

Let E and F be Banach spaces, U an open subset of E, and f a C^1 mapping from U to F. Then Df is a continuous mapping from U to $\mathcal{L}(E; F)$. But this Df is a mapping from an open subset (U) of a Banach space to a Banach space. If this mapping is in fact differentiable, and if its derivative, DDf is continuous, then f is said to be twice continuously differentiable, or C^2. In this case, DDf is a continuous mapping from U to $\mathcal{L}(E; \mathcal{L}(E; F))$. Similarly, if this mapping is C^1 (so $DDDf$ exists as a continuous mapping from U to $\mathcal{L}(E; \mathcal{L}(E; \mathcal{L}(E; : F))))$, then f is said to be C^3. Similarly for C^p, p a non-negative integer. Finally, f is said to be C^∞ if f is C^p for every p.

Example. The mapping of the first example on page 17 is C^∞. Since Df is constant ($Df(x_0) = T$ for every x_0), $DDf = 0$ $DDDf = 0$, etc.

Example. The sum of two C^p mappings is C^p. This is immediate from the facts that the sum of two C^1 mappings is C^1, that the derivative of the sum is the sum of the derivatives.

Example. The mapping of the first example on page 18 (beginning on the previous page) is C^∞. Indeed, DDf is that element of DDf is that element of $\mathcal{L}(E; \mathcal{L}(E; P))$ which sends φ in E to the mapping from E to P which sends φ' to the element $2\varphi\varphi'$ of F. Hence, DDf is constant. So, $DDDf = 0$, etc. [Just what one expects of a "quadratic mapping".]

Example. Let E and F be Banach spaces, and T an element of $\mathcal{L}(E, \ldots, E; F)$. Set $f(x) = T(x, \ldots, x)$. Then f is C^∞.

We next show that composition of C^p mapping is C^p.

Theorem. Let E, F and G be Banach spaces, U an open subset of E and V an open subset of F. Let f be a C^p mapping from U to F with $f(U) \subset V$ and g a C^p mapping from V to G. Then the mapping $g \cdot f$ from U to G is also C^p.

Proof: suppose first that f and g are C^0. Then, given x_0 in U and positive ϵ, first choose δ' such that $|g(y) - g(f(x_0))| \leq \epsilon$ whenever $|x - x_0| \leq \delta'$. Then choose δ, such that $|f(x) - f(x_0)| \leq \delta'$ whenever $|x - x_0| \leq \delta$. Then, for $|x - x_0| \leq \delta$, we have $|g \cdot f(x) - g \cdot f(X_0)| \leq \epsilon$. That is, $g \cdot f$ is C^0.

Next, suppose that f and g are C^1, Fix x_0 in U. Then $Dg(f(x_0))$ is an element of $\mathcal{L}(F; G)$, while $Df(x_0)$ is an element of $\mathcal{L}(E; F)$. Their composition is thus an element of $\mathcal{L}(E; G)$. Note first that the mapping with action $f(x) - f(x_0) - Df(x_0)(x - x_0)$ is tangent at $x = x_0$. Applying $Dg(f(x_0))$ to this expression, using the fact that the composition of a bounded linear mapping with a function tangent is a function tangent, we have that $Dg(f(x_0))(f(x) - f(x_0)) - Dg(f(x_0))Df(x_0)(x - x_0)$ is tangent at $x = x_0$. But, since $g(y) - g(f(x_0)) - Dg(f(x_0))(y - f(x_0))$ is tangent at $y = f(x_0)$, and since f is continuous, $g(f(x)) - g(f(x_0)) - Dg(f(x_0))(f(x) - f(x_0))$ is tangent at $x = x_0$. Adding these two, we obtain: $g(f(x)) - g(f(x_0)) - Dg(f(x_0))Df(x_0)(x - x_0)$ is tangent at $x = x_0$. But this is precisely the statement that $g \cdot f$ is differentiable at x_0, with $D(g \cdot f)(x_0) = Dg(f(x_0))Df(x_0)$. Next note that $f(x)$ is continuous in x and $Dg(y)$ is continuous in y, whence their composition, $Dg(f(x))$ is continuous in x. Further, $Df(x)$ is continuous in x. But the element of $\mathcal{L}(\mathcal{L}(E; F), \mathcal{L}(F; G); \mathcal{L}(E; G))$ which composes is also continuous (pp 16). Hence, $Dg(f(x))Df(x)$ is a continuous mapping from U to $\mathcal{L}(E; G)$. We conclude that $g \cdot f$ is C^1.

Now suppose that both f and g are C^2. Since $f(x)$ is C^1 in x, and $Dg(y)$ is C^1 in y, $Dg(f(x))$ is C^1 in x (last paragraph). Also, $Df(x)$ is C^1 in x. But composition is also C^1 (since it is bilinear). Hence, $Dg(f(x))\ Df(x)$ is C^1 in x. That is, $D(g \cdot f)$ is C^1, whence $g \cdot f$ is C^2. Continue in the obvious way to C^p.

We next show that operation "taking the inverse" is also C^∞.

Example. Let E and F be Banach spaces. Denote by U the subset of $\mathcal{L}(E; F)$ consisting of all T therein which are invertible. We first show that this subset U is open. Fix T_0 in U. Then T_0 is an isomorphism from E to F. Consider the isomorphism from $\mathcal{L}(E; F)$ to $\mathcal{L}(E; E)$ which sends T in $\mathcal{L}(E; F)$ to $T\,{T_0}^{-1}$. Under this isomorphism, T_0 itself is sent to I, the identity on E. It suffices, therefore, to show that there is some ball centered at I in $\mathcal{L}(E; E)$ every element of which is invertible. Let B be the ball with radius $1/2$. Let W be an element of B. Consider the sequence $V - 1 = I, V - 2 = I + (I - W)$, $V_3 = I + (I - W) + (I - W)(I - W)$, etc. This is a Cauchy sequence in $\mathcal{L}(E; E)$, for $|V_{i+1} - V_i| = |(I - W)^i| \leq |I - W|^i \leq 1/2^i$. Hence, this sequence converges to some element V of $\mathcal{L}(E; E)$. We now have, by direct computation, $WV_i - I = V_i - V_{i+1}$. Taking the limit of each side as i increases, noting that that on the right is zero, we obtain $WV = I$, and, similarly, $VW = I$. Thus, V is the inverse of W, and so W is invertible. We have shown so far that our subset U of $\mathcal{L}(E; F)$ is open. Note also that, above, $|W^{-1}| \leq 1 + 1/2 + 1/4 + \ldots = 2$.

Next, introduce the mapping f from U to $\mathcal{L}(F; E)$ with the following action: For T in U, $f(T) = T^{-1}$. We next show that this mapping f is continuous. Fixing T_0 in U, we have for T in U, $f(T) - f(T_0) = T^{-1} - {T_0}^{-1} = -T^{-1}(T - T_0){T_0}^{-1}$, whence $|f(T) - f(T_0)| \leq |T^{-1}||T - T_0||{T_0}^{-1}|$. Given positive ϵ, choose δ such that, whenever $|T - T_0| \leq \delta$, $|T^{-1}| \leq M$, some constant (possible, from the observation above that, whenever $|I - W| \leq 1/2$, $|W^{-1}| \leq 2$), and such that $\delta \leq \epsilon / M{T_0}^{-1}$. Then, for $|T - T_0| \leq \delta$, we have $|f(T) - f(T_0)| \leq \epsilon$. Thus, f is continuous at T_0, which is arbitrary in U, and so f is continuous.

We next show that this mapping f is in fact differentiable at $T = T_0$. Its derivative at T_0 must be an element of $\mathcal{L}(\mathcal{L}(E; F); \mathcal{L}(F; E))$. From elementary calculus, we guess the derivative as follows: Let α be the bounded linear mapping from $\mathcal{L}(E; F)$ to $\mathcal{L}(F; E)$ which sends T in $\mathcal{L}(E; F)$ to $\alpha(T) = -{T_0}^{-1}T{T_0}^{-1}$. Then $f(T) - f(T_0) - \alpha(T - T_0) = T_{-1}^{-1} - {T_0}^{-1} + {T_0}^{-1}(T - T_0){T_0}^{-1} = T^{-1}(T - T_0)T_0(T - T_0){T_0}^{-1}$. Hence, we have $|f(T) - f(T_0) - \alpha(T - T_0)| \leq |T^{-1}||T_0|^{-1/2}|T - T_0|^2$. Now, given positive ϵ, choose δ such that,

whenever $|T - T_0| \leq \delta$, we have $|T^{-1}| \leq M$ (some constant), and such that $\delta \leq \epsilon/M|T_0{}^{-1}|^2$. Then, for $|T - T_0| \leq \delta$, we have $|f(T) - f(T_0) - \alpha(T - T_0)| \leq \epsilon|T - T_0|$. We conclude that the mapping f from U to $\mathcal{L}(F; E)$ is indeed differentiable at T_0 in U, and that its derivative is α.

For each T_0 in U, we obtain, as above, the derivative of f at T_0, $Df(T_0)$. Since "take the inverse" and "compose" are continuous operations, this Df is a continuous mapping from U to $\mathcal{L}(F; E)$. So, f is C^1. But now, since "take the inverse" and "compose" are C^1 operations, Df is C^1, whence f is C^2. etc. So f is C^p.

All of the little calculations above are identical to those one is familiar with in the finite-dimensional case.

Finally, we establish the result in the present context which generalizes the finite-dimensional statement that "mixed partials commute". For this purpose, we note that DDf maps U to $\mathcal{L}(E; \mathcal{L}(E; F))$, and that $(E; F))$ is isomorphic with the Banach space $\mathcal{L}(E, E; F)$. Hence, we may regard $DDf(x_0)$ as an element of $\mathcal{L}(E, E; F)$.

Theorem. Let E and F be Banach spaces, U open is E, and f a C^2 mapping from U to F. Then DDf is symmetric: For x_0 in U and x, y in E, $DDf(x_0)(x, y) = DDf(x_0)(y, x)$.

Proof : Fix an element α of $\mathcal{L}(F; \mathbb{R})$. Let h be the function of two real variables a and b with action $h(a, b) = \alpha(f(x_0 + ax + by))$. [Thus, h is only defined for a and b small enough that $x_0 + ax = by$ is in U.] Using the chain rule (preceding theorem), we have $\partial h/\partial a = \alpha(Df(x_0 + ax + by)(x))$, and thus $\partial^2 h/\partial a\,\partial b = \alpha(DDf(x_0 + ax + by)(y, x))$. Evaluating at $a = b = 0$, we have $\partial^2 h/\partial a\,\partial b|_{a=b=0} = \alpha(DDf(x_0)(y, x))$, and similarly $\partial^2 h/\partial b\,\partial a|_{a=b=0} = \alpha(DDf(x_0)(x, y))$. But the left sides are equal (equality of mixed partials for real functions of two real variables), and so $\alpha(DDf(x_0)(y, x) - DDf(x_0)(x, y) = 0$ for every α. Thus, we shall be done if we can show that the only element z of F such that $\alpha(z) = 0$ for every α in $\mathcal{L}(F; \mathbb{R})$ is $z = 0$. Given z in F, choose complementary subspace G in F to the subspace spanned by z. Then any vector u in F can be written uniquely in the form $u = az + v$ with v in G. Let α have action $\alpha(u) = a$. Then α is bounded linear mapping (obviously linear, and bounded since α is bounded on both G and the subspace spanned by z, and since F is isomorphic with the product of these, by the open mapping theorem). but $\alpha(z) = 1$. So, this α is the thing we wanted to find. This completes the proof.

The proof above is rather tacky, because it uses itself in the finite-dimensional case, which is taken as known. Unfortunately, the proof of this theorem in finite-dimensions does not seem to be directly generalizable to infinite dimensions. I am aware of no more direct proof of the theorem above.

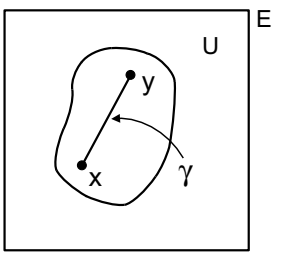
It follows immediately by repeated application of this result that higher mixed partials are also symmetric. For example, $DDDf(x_0)$, an element of $\mathcal{L}(E; \mathcal{L}(E; \mathcal{L}(E; F)))$, can also be regarded as an element of $\mathcal{L}(E, E, E; F)$. In this case we have, for x, y, and z in E and for fC^3, that $DDDf(x_0)(x, y, z) = DDDf(x_0)(y, x, z) = DDDf(x_0)(z, y, x)$, etc.

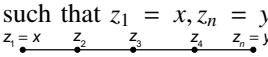
8. Mean Value Theorem; Inverse Mapping Theorem

We complete our treatment of elementary calculus in infinite dimensions by proving the two titled theorems.

The mean value theorem states that the derivative of a function, a measure of its rate of change, in fact bounds the actual changes in the value of the function.

Theorem (mean value). Let E and F be Banach spaces, U an open subset of E, and f a C^1 mapping from U to F. Let x and y be points of U such that the line segment γ joining x and y lies within U. Then $|f(y) - f(x)| \leq \operatorname{lub}_{z \text{ in } \gamma} |Df(x)||y - x|$.



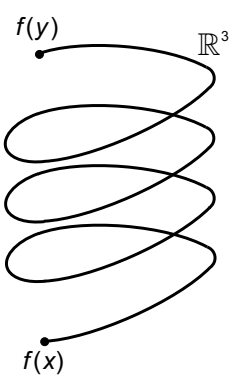
Proof: [Of course, γ is the set of vectors of the form $ax+(1-a)y$, with $0 \leq a \leq 1$.] Fix, once and for all, positive number ϵ. Given z_0 on γ, the function with action $f(z) - f(z_0) - Df(z_0)(z - z_0)$ is tangent at $z = z_0$. Hence, there is a positive δ such that, whenever $|z - z_0| \leq \delta$, we have $|f(z) - f(z_0)| \leq (|Df(z_0)| + \epsilon)|z - z_0|$. Thus, for every z_0 on γ we have such a δ. Since γ is compact, we can find points $z_1, \ldots, z_n$ on γ such that $z_1 = x, z_n = y, z_i \leq z_{i+1}$, (in the obvious sense: 

), and such that $|z_i - z_{i+1}|$ is less than the δ appropriate to z_{i+1}. Now, we have $|f(y)-f(x)| \leq |f(z_n)-f(z_1)|\ |f(z_n)-f(z_{n-1})|+ \ldots+|f(z_2)-f(z_1)|$. But $|f(z_{i+1})-f(z_i)| \leq |(Df(z_i)| = \epsilon)|z_{i+1}-z_i| \leq (\text{lub}\,|\text{Df(z)}|+\epsilon)|\text{z}_{\text{i+1}}-\text{z}_\text{i}|$. Hence, $|f(y) - f(x)| \leq (\text{lub}\,|\text{Df(z)}| + \epsilon)(|\text{z}_\text{n} - \text{z}_{\text{n-1}}| + \ldots + |\text{z}_2 - \text{z}_1|)$. But the sum on the right is just $|z_n - z_1| = |y - x|$. So, we have $|f(y) - f(x)| \leq (\text{lub}\,|\text{Df(z)}| + \epsilon)|\text{y} - \text{x}|$. Since, ϵ is arbitrary, the result follows.

One might imagine that another version of the mean value theorem, which on the one hand would be stronger and on the other would be more closely analogous to the one-dimensional mean value theorem, might be true: Under the conditions of the theorem above, there is a point z_0 of γ such that $f(y)-f(x) = Df(z)(y-x)$. This, however, is false.

Example. Let E be the reals, and $F = R^3$. Then a C^1 mapping f from $U = E$ to F represents a curve in $\mathbb{R}^3$. $Df(z)$ is the tangent vector to this curve at z in R.

On the other hand, $f(y)-f(x)$, is the vector joining these two points in $\mathbb{R}^3$. Thus, we must find a curve joining two points in $\mathbb{R}^3$, whose tangent vector is never parallel to the vector joining those points (for then we shall never have $f(y) - f(x) = Df(z)(y - x)$). But a spiral, as in the figure above, does the job.



Example. Let $Df = 0$. Then, by the mean value theorem, f is constant (i.e., $f(x) = f(y)$ for all x, y such that the line segment joining them lies in U). Similarly, if $DDf = 0$, then the action of f is $f(x) = T(x) + x_0$, for some x_0 in E and T in $\mathcal{L}(E; F)$ (for Df is some constant T, and then $D(f - T(x)) = 0$, whence this is some constant x_0.)

The inverse function theorem states that, if a mapping is "invertible to first order some point, then it is actually invertible in some ball about that point". As a prerequisite to this theorem, we need a little fact about mappings (a fact which, indeed, is the basis for all existence and most uniqueness theorems about differential equations). Let C be a closed subset of a Banach space, and f a mapping from C to C such that, for some number $a < 1$, $|f(x) - f(y)| \leq a|x - y|$ for every x, y in C. Then there is one and only one x in C with $f(x) = x$ Proof: First note that, if there were two, x and y, then we would have $|y - x| = |f(y) - f(x)| \leq a|y - x|$, whence, since $a < 1, |y - x| = 0$, so $y = x$. Thus, we have only to show that $f(x) = x$ for some x. Choose y in C, and let

$y_1 = y, y_2 = f(y_1), y_3 = f(y_2)$, etc. Then, since $|y_{i+1} - y_i| = |f(y_i) - f(y_{i-1})| \leq a|y_i - y_{i-1}|$, we have $|y_i - y_j| \leq a^j + \ldots + a^{j-1}$ (for $j > i$), whence the y_i form a Cauchy sequence. Since the y_i are in a Banach space, this sequence converges to some vector x; since C is closed and the y_i are in C, x is in C. But now $f(y_i) - x = y_{i+1} - x$. Taking the limit of each side as x increases, noting that that on the right is zero, we obtain $f(x) = x$.

Theorem (inverse mapping). Let E and F be Banach spaces, U open in E. and f a C^p mapping from U to F, with $p \geq 1$. Let, for some x_0 in U. the element $Df(x_0)$ of $\mathcal{L}(E; F)$ be invertible. Then there is some open subset V of U containing x_0 such that i) $f[V]$ is open in F, ii) f is one-to-one on V, and iii) the inverse mapping f from $f[V]$ to E is also C^p.

Proof: Let α be the mapping from F to E with action $\alpha(y) = (Df(x_0)^{-1}(y - f(x_0)) + x_0$. Then α is C^∞. Furthermore, the C^∞ β from E to F with action $\beta(x) = Df(x_0)(x - x_0) + f(x_0)$ is its inverse, i.e., $\alpha \cdot \beta$ is the identity on E and $\beta \cdot \alpha$ is the identity on F. We may consider, instead of f, $\alpha \cdot f$. that is to say, we may assume that f maps U to E, that $x_0 = 0$, that $f(x_0) = 0$, and that $Df(0)$ is the identity element of $\mathcal{L}(E; E)$.

Let g be the mapping from U to E which sends x in U to $g(x) = x - f(x)$. Then $g(0) = 0$, and $Dg(0) = D$ (identity) $(0) - Df(0) = 0$. Since g is C^1, there is a positive ϵ such that, whenever $|x| < 2\epsilon$, $|Dg(x) < 1/2$. Then, from the mean value theorem, we have, for $|x| < 2\epsilon$, $|g(x)| < 1/2\ |x|$.

Next, fix y in E with $|y| \leq \epsilon/2$. Denote by g_y the mapping from U to E with action $g_y(x) = y + x - f(x)$. Then, for $|x| \leq \epsilon$ we have $|g_y(x)| = |y + g(x)| \leq |y| + |g(x)| \leq 1/2\epsilon + 1/2|x| \leq \epsilon$. Thus, denoting by B the set of all vectors x with $|x| \leq \epsilon$, we have that g_y is a mapping from B to B. Furthermore, for x and x' in B, we have $|g_y(x) - g_y(x')| = |g(x) - g(x')| \leq 1/2\ |x - x'|$, again by the mean value theorem (and the fact that $|Dg(x)| < 1/2$ for $|x| < 2$). Thus, g_y, as a mapping from the closed subset B of E to itself, a mapping with the contraction property of the previous page. Hence, there is one and only one x in B with $g_y(x) = x$, i.e., with $y + x - f(x) = x$, i.e., with $y = f(x)$.

Now let V consist of all x in U with $|x| < \epsilon$ and $|f(x)| < \epsilon/2$. Then V (as the intersection of U, an open ball, and the inverse image by f of an open ball) is open. Furthermore, $f[V]$ is just the set of y with $|y| < \epsilon/2$, and so is open in E. In addition, f is one-to-one on V (for, as we showed above, given y with $|y| \leq \epsilon/2$, there is one and only one x with $|x| \leq \epsilon$ and $f(x) = y$)

Denote by $\hat{f}$ the inverse of f, defined on $W = f[V]$. Then, for x and x' in V, $|x - x'| = |f(x) + g(x) - f(x') - g(x')| \leq |f(x) - f(x')| + |g(x) - g(x')| \leq |f(x) - f(x')| + 1/2|x - x'|$, whence $|x - x'| \leq 2|f(x) - f(x')|$. Setting $y = f(x)$ and $y' = f(x')$, this becomes $|\hat{f}(y) - \hat{f}(y')| \leq 2|y - y'|$. That is to say, this f is continuous.

With x, x', y and y' as above, we have $|\hat{f}(y) - \hat{f}(y') - (Df(x'))^{-1}(y - y')| = |(Df(x'))^{-1}(Df(x')(x - x') - f(x) + f(x'))| \leq |(Df(x'))^{-1}|\,|Df(x')(x - x') - f(x) + f(x')|$. Fixing x', the right side, regarded as a function of x, is tangent at x', since $\hat{f}$ is differentiable. Hence, by continuity of $\hat{f}$, the left side is tangent at $y = y'$. We conclude, therefore, that $\hat{f}$ is differentiable, and that $D\hat{f}(y) = (Df(\hat{f}(y))^{-1}$. Since $\hat{f}$ is continuous, Df is continuous, composition is continuous, and "take the inverse" is continuous, $D\hat{f}$ is continuous. That is, $\hat{f}$ is C^1.

If f is C^2, then, since $\hat{f}$ is C^1, Df is C^1, composition is C^1, and "take the inverse" is C^1, $D\hat{f}$ is C^1, and so $\hat{f}$ is C^2. Continuing in this way, if f is C^p, then so is $\hat{f}$.

There is a corollary of the inverse function theorem, called the implicit function theorem. It states that one can, under certain circumstances, "solve" equations of the form $f(x, y) = 0$ for x as a function of y. More precisely, let E and F be Banach spaces, and let U and V be open subsets of E and F, respectively. For each y in V, let f_y be a mapping from U to G. [We may thus regard f as a mapping from $U \times V$ to G; above, we write $f(x, y)$ for $f_y(x)$.] Suppose that i) $_y(x)$ is continuous on $U \times V$, ii) for each y in V, f_y is C^1, and $Df_y(x)$ is continuous on $U \times V$, and iii) for some x_0 in U and y_0 in V, we have $f_{y_0}(x_0) = 0$ and $Df_{y_0}(x_0)$, an element of $\mathcal{L}(E; G)$, is invertible. Then there is an open subset V_0 of V containing y_0, and a mapping g from V_0 to U, such that $f_y(g(y)) = 0$ for every y in V_0. [Thus, $x = g(y)$ is the "solution" of $f_y(x) = 0$.] Proof: Set $h_y(x) = f_y(x) - f_y(x_0)$.

Then $h_y(x_0) = 0$ for every y. Furthermore, $Dh_y = Df_y$. Hence, since $Df_{y_0}(x_0)$ is invertible, since $Df_y(x)$ is continuous, and since the invertible elements of $\mathcal{L}(E;G)$ form an open subset of $\mathcal{L}(E;G)$, we have that $Dh_y(x-0)$ is invertible for all y sufficiently close to y_0. Applying the inverse function theorem to h_y, for each y, we have: There is an open subset W of G containing 0, and a mapping s_y from W to E, for y sufficiently close to y_0, such that $h_y \cdot s_y$ is the identity on W. [We may choose W independent of y because the size of the "V" in the inverse function theorem is dictated by the size of "$'Df$" in that theorem, and here $Df_y(x)$ is continuous.] Since $f_y(x)$ is continuous in y and x, it follows that, for y sufficiently close to y_0, $f_y(x_0)$ is in this W. Now set $g(y) = s_y(f_y(x_0))$. Then, since $h_y \cdots_y$ = identity, $h_y(g(y)) = f_y(x_0)$. That is to say, $f_y(g(y)) - f_y(x_0) = -f_y(x_0)$, or $f_y(g(y)) = 0$.

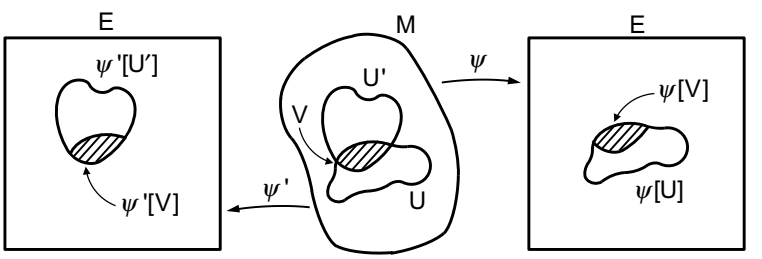
Note that in the above E and G must be isomorphic as Banach spaces. This requires, essentially, that the "number of equations" (represented by the "size" of G) be the same as the "number of unknowns" (represented by the size of E).

9. Manifolds

We have now completed our discussion of differential calculus on Banach spaces. We turn next to that mathematical object – a manifold – which be at the center of all we do hereafter. In this section, we define manifolds, give some examples, and give some constructions which yield manifolds.

A manifold is, roughly speaking, a space having the "local smoothness structure" of some Banach space. The idea, then, is to isolate, from the very rich structure of a Banach space (e.g., its vector-space structure, its norm structure) that one type of structure we call "local smoothness".

We first introduce a mechanism by which structure can be carried from Banach spaces to other things. Fix a set M and a Banach space E. A *chart* (or $E-$chart) on M consists of a subset U of M together with a mapping ψ from U to E, such that i) the mapping ψ is one-to-one, and ii) the subset $\psi[U]$ of E, the image of U by ψ, is open in E. Thus, a chart sets up a one-to-one correspondence between a certain subset, U, of M and a certain open subset, $\psi[U]$, of E. It is by means of this correspondence that structure is carried from E to M.



We next introduce a notion of "agreement between two charts as regards their induced smoothness structures on M". Let U, ψ and U', ψ' be two E-charts on the set M. Then on the intersection of their U's, $V = U \cap U'$, there are induced two "smoothness structures", one from ψ (which defines a correspondence between V and the subset $\psi[V]$ of E) and the other from ψ' (which defines a correspondence between V and the subset $\psi'[V]$ of E). We wish to compare these.

To this and, introduce the mapping $\psi' \cdot \psi^{-1}$ from $\psi[V]$ to $\psi'[V]$, and its inverse, $\psi \cdot \psi'^{-1}$, from $\psi'[V]$ to $\psi[V]$. These mapping represent the interaction between U, ψ, and U', ψ'. [Note that M has now been eliminated: The mappings are between subsets of Banach spaces.]

Now fix a symbol p, either a non-negative integer or the symbol "∞". We are led to the following definition: The charts U, ψ and U', ψ' on M are said to be *compatible* (or C^p-compatible) if i) $\psi[V]$ and $\psi'[V]$ are both open subsets of E, and ii) the mapping $\psi' \cdot \psi^{-1}$ from $\psi[V]$ to E, and $\psi \cdot \psi'^{-1}$, from $\psi'[V]$ to E, are both C^p mappings. The second condition is the crucial one. Note that we do not require that our mappings preserve vector-space structure, or norm structure. Rather, they need only preserve C^p differential structure. It is in this way that a single "type of structure" is isolated. Note also that two charts are necessarily compatible if their U's a fail to intersect.

Example. Let U, ψ be an $E-$chart on M, let U' be a subset of U such that $\psi[U']$ is an open subset of $\psi[U]$, and let ψ' be ψ restricted to U'. Then U, ψ and U', ψ' are compatible.

A *manifold* consists of a non-empty set M, a Banach space E, a symbol p, and a collection ζ of $E-$charts on M, satisfying the following conditions:

1. Any two charts in the collection ζ are C^p compatible.
2. The charts in ζ cover M, i.e., every point of M is in at least one of the $U's$.
3. Any chart on M which is compatible with all the charts in ζ is itself in ζ.

4. The charts separate points of M in the following sense: Given distinct points p and p' of M, there are charts U, ψ and U', ψ' in ζ such that p is in U and p' in U', and such that there is a ball B centered at $\psi(p)$ in $\psi[U]$ and a ball B' centered at $\psi'(p')$ in $\psi'[U']$, with $\psi^{-1}[B]$ and $\psi' - 1[B']$ not intersecting in M.

These conditions – or at least the first three – are exactly what one might have expected intuitively. The first requires that "whenever two charts in ζ induce smoothness structures in the same region of M, these structures agree". The second requires that "smoothness structure has been induced over all of M". The third guarantees that we have not induced any additional structure on M by cutting down the number of charts. Finally, the fourth condition (which is normally automatically satisfied in practice) eliminates certain pathological object, called non-Hausdorff manifolds, which are of little interest.

The manifold defined above is sometimes called a C^p manifold based on Banach space E. The charts in ζ are called the admissible charts, or just the charts. We often denote a manifold just "M", the rest understood. A subset O of manifold M is said to be *open* if, for every admissible chart U, ψ, $\psi[U \cap O]$ is open in E. These are of course, the open sets for a topology on M. The fourth condition then requires that this topological space be Hausdorff. Our third condition is often omitted, and our fourth occasionally.

One might imagine that it would be very difficult to give any examples of manifolds, for, by the third condition above, there will be an enormous collection of charts, and it might be awkward to write all these down. The possibility of giving examples easily arises from the following fact.

Lemma. Let M be a non-empty set, E a Banach space, p a symbol (a non-negative integer or "∞"), and $\hat{\zeta}$ a collection of $E-$charts on M satisfying conditions one, two, and four above. Denote by ζ the collection of all charts on M compatible with all those in $\hat{\zeta}$. Then ζ satisfies all four conditions.

Proof: First note that, since every chart in $\hat{\zeta}$ is necessarily in ζ, this ζ automatically satisfies the second and third conditions. Similarly, the fourth condition is also automatic. Thus, we need only verify the first. Let U, ψ and U', ψ' be two charts in ζ, and set $V = U \cap U'$. Fix point p of V, and choose chart $\hat{U}, \hat{\psi}$ in $\hat{\zeta}$, with p in $\hat{U}$ (possible, by second condition). By compatibility of $\hat{U}, \hat{\psi}$ and $U\psi$, and of $\hat{U}, \hat{\psi}$ and U', ψ', there is a ball B centered at $\hat{\psi}(p)$ such that B is in $\hat{\psi}[V]$ $(= \hat{\psi}[U]\hat{\psi}[U'])$. By compatibility of $\hat{U}, \hat{\psi}$ and U, ψ, $\psi \cdot \hat{\psi}^{-1}[B]$, a subset of $\hat{\psi}[V]$, contains a ball centered at $\psi(p)$. Thus, $\psi[V]$ is open in E, and similarly for $\psi'[V]$. Finally, since $\psi \cdot \hat{\psi}^{-1}$ and $\hat{\psi} \cdot \psi'^{-1}$ are C^p, so is their composition, $\psi \cdot \psi'^{-1}$ and similarly for $\psi' \cdot \psi^{-1}$. We conclude that U, ψ and U', ψ' are compatible.

Thus, to obtain a manifold, one need only find charts satisfying the first, second, and fourth conditions: Something which is often rather easy to do.

Examples. Let M be a C^p manifold based on E, and let $q \leq p$ (where, of course, the non-negative integers are ordered in the usual way, $q \leq \infty$ for every q, and $\infty \leq p$ only if $p = \infty$). Then, since charts C^p-compatible are also C^q-compatible, the admissible charts on M satisfy the first, second, and fourth condition with p replaced by q. By the Lemma, we obtain a C^q manifold based on E. The question of whether, given a C^q manifold and $p > q$, one can throw away some charts to get a C^p manifold is very difficult.

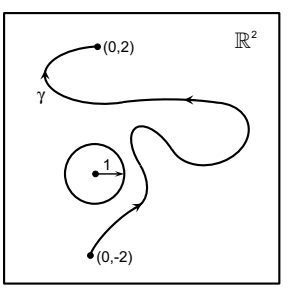
Example. Let E be a Banach space, and let M be the set E. Introduce a chart with $U = M$ (a subset of M), and ψ the identity mapping (from $U = M = E$ to E). This single chart satisfies the first condition, since every chart is compatible with itself, and obviously the second. This chart also satisfies the fourth condition, since given distinct points of a Banach space, one can find non-intersecting balls centered at those points. By the Lemma, we obtain a C^∞ manifold based on E.

Example. Let M denote the set of all sequences, $(r_1, r_2, \ldots)$, of real numbers with $(r_1)^2 + (r_2)^2 + \ldots = 1$. Let E be the Banach space of all sequences $(s_1, s_2, \ldots)$ the sum of the squares of whose entries converges. We introduce some $E-$charts on M. Let U be the subset of M consisting of all elements with $r_1 > 0$. For $(r_1, r_2, \ldots)$ in U, set $\psi(r_1, \ldots) =$

$(r_2, r_3, \ldots)$ (an element of E, since the sum of the squares of the entries converges). We claim that this is a chart. Indeed, for $\psi(r_1, r_2, \ldots) = \psi(r'_1, r'_2, \ldots)$, we must certainly have $r_2 = r'_@$, $r_3 = r'_3$, etc. But we must also have $r_1 = r'_1$, for each is the positive square root of one minus the sum of the squares of the other entries. That is, ψ is one-to-one. We next claim that $\psi[U]$ is the subset of E consisting all $(s_1, s_2, \ldots)$ with $(s_1)^2 + \ldots < 1$, a claim which is obvious from the observation that $\psi(\sqrt{1 - \Sigma(s_i)^2}, s_1, s_2, \ldots) = (s_1, s_2, \ldots)$. But this $\psi[U]$ is an open subset of E (in fact, is the ball of radius one centered at the origin). Thus, we have a chart.

Similarly, we obtain a chart with U consisting of $(r_1, r_2, \ldots)$ with $r_1 < 0$, and the same ψ. Doing the same thing with r_2, and then with r_3, etc., we obtain still more charts on M.

We next claim that these charts satisfy conditions one, two, and four, for $p = \infty$. The second and fourth are immediate, and so we have only to check compatibility. Let us consider, e.g., the chart U, ψ (where U requires $r_1 > 0$) and U', ψ' (where U' requires $r_2 > 0$). Then $V = U \cap U'$ consists of $(r_1, r_2, \ldots)$ in M with $r_1 > 0$ and $r_2 > 0$, whence $\psi[V]$ consists of $(s_1, s_2, \ldots)$ with $(s_1)^2 + \ldots < 1$ and with $s_1 > 0$. But this is certainly an open subset of E. Similarly for $\psi'[V]$. Now, for $(s_1, s_2, \ldots)$ in $\psi'[V]$, we have $\psi^{-1}(s_1, \ldots) = (\sqrt{1 - \Sigma(s_i)^2}, s_1, s_2, \ldots)$, whence $\psi' \cdot \psi^{-1}(s_1, \ldots) = (\sqrt{1 - \Sigma(s_i)^2}, s_2, s_3, \ldots)$. We claim that this mapping $\psi' \cdot \psi^{-1}$ from $\psi[V]$ to E is C^∞. It is derivative at $(s_1, s_2, \ldots)$ in $\psi[V]$, for example, sends $(t_1, t_2, \ldots)$ in E to the element $(\theta(1 - \Sigma(s_i)^2)^{-1/2}(s_1 t_1 + s_2 t_2 + \ldots)$, $t_2, t_3, \ldots)$ of E. [This guess is made, of course, from ordinary calculus.]



Thus, our charts satisfy the first, second, and fourth conditions. By the Lemma, we obtain a manifold. [This is the infinite-dimensional sphere. Similarly, one obtains the finite-dimensional sphere.]

Example. Let F be the Banach space of continuous, real-valued functions f on the closed interval $[0, 1]$ with $f(0) = f(1) = 0$. Let $E = F \times F$, product of Banach spaces. Next, denote by M the set of all curves in the plane which begin at $(0, -2)$, end at $(0, +2)$, and which avoid the closed disk of radius one centered at the origin. [That is, M is the set of all continuous maps γ from $[0, 1]$ to $\mathbb{R}^2$ with $\gamma(0) = (0, -2), \gamma(1) = (0, +2)$, and with $|\gamma(r)| \leq 1$ for no r.] We introduce an $E-$chart on this set M. Fix an element $\gamma(0)$ of M. Let $U = M$, and let ψ be the mapping from U to E which sends γ in M to the element $\gamma(r) - \gamma_0(r)$ of E (noting that $\gamma - \gamma_0$ as a continuous map from $[0, 1]$ to $\mathbb{R}^2$ beginning and ending at the origin, is an element of E). This ψ is clearly one-to-one. Furthermore, $\psi[U]$ is open in E, for, given $\psi(\gamma)$ in E, one can find an ϵ such that, whenever $|\psi(\gamma) - \tau| \leq \epsilon$, τ is also in $\psi[U]$. Thus, this is a chart.

Again, this single chart satisfies the first, second, and fourth conditions for a manifold. By the Lemma, we obtain a manifold.

We shall later display some more interesting examples of manifolds. There are two elementary techniques for constructing new manifolds from old. These techniques are of interest, first because they give some insight into what a manifold is, second because they yield a rich source of examples, and third because the techniques themselves often arise in practice. We now discuss these techniques.

Let M be a manifold based on Banach space E, and let N be a subset of M and F a subspace of E. We suppose that these objects are related as follows: Given any point p of N, there is an admissible chart U, ψ on M such that p is in U, and such that $\psi[U \cap N] = \psi[U] \cap F$. Of course, such a

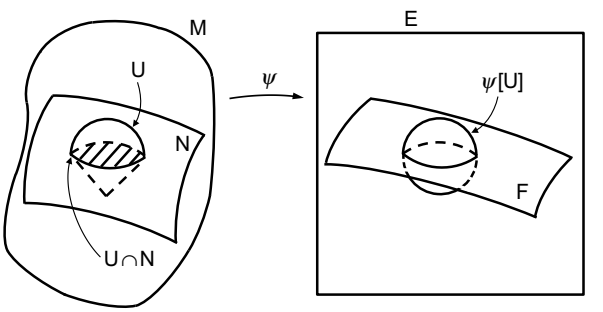


chart will not in general exist: our requirement of its existence represents an additional condition on N and F. this condition requires essentially that "N in M look like F in E".

We next introduce some $F-$charts on this set N. Let U, ψ be an admissible chart on M, with $\psi[U \cap N] = \psi[U] \cap F$. Set $W = U \cap N$, and let φ be the restriction of ψ to W. Then W is certainly a subset of N, and φ certainly maps this W to Banach space F. Furthermore, since $\varphi[W] = \psi[U] \cap F$, this $\varphi[W]$ is an open subset of F, while, since ψ is one-to-one, so is φ. We conclude that W, φ is an $F-$chart on N.

The $F-$charts on N, so obtained, clearly satisfy the fourth condition for a manifold (since the admissible charts on M do). The second condition follows from the condition above (since the above is precisely the statement that our charts cover N). We next verify the first condition, i.e., we show that any two such charts are compatible. Let W, φ and W', φ be two $F-$charts on N, obtained from charts U, ψ and U', ψ' on M, respectively. Then, since $\varphi[W \cap W'] = \psi[U \cap U'] \cap F$, and since the right side is an open subset of F, so is the left side. Similarly, $\varphi'[W \cap W']$ is an open subset of F. Thus, there remains only to show that the mapping $\varphi' \cdot \varphi^{-1}$ from $\varphi[W \cap W']$ to F is C^p. To this end, fix x_0 in $\varphi[W \cap W']$, and write α for $\varphi' \cdot \varphi^{-1}$ and β for $\psi' \cdot \psi^{-1}$. Then, since β is differentiable at x_0, we have that $\beta(x) - \beta(x_0) - D\beta(x_0)(x - x_0)$ (as x ranges over points of E sufficiently close to x_0) is tangent at $x = x_0$). Now fix x in F: Then $(x - x_0)$, $\beta(x)$, and $\beta(x_0)$ are all in F, We claim that, therefore, $D\beta(x_0)(x - x_0)$ must also be in F. [Proof: Suppose not. Then there is a positive ϵ such that $|D\beta(x_0)(x - x_0) - y| \geq \epsilon$ for every y in F. Then, for all positive a, we have $|\beta(ax + (1 - a)x_0) - \beta(x_0) - D\beta(x_0)(ax - ax_0)| \geq a\epsilon$, since the first two terms are in F. But, setting, $x' = ax + (1 - a)x_0$ and letting a go to zero, this last inequality violates tangency of $\beta(x') - \beta(x_0) - D\beta(x_0)(x' - x_0)$ at $x' = x_0$.]. Denote by T the mapping $D\beta(x_0)$ restricted to F, so T is a bounded linear mapping from F to F. Then we have, for x in F, $\beta(x) - \beta(x_0) - D\beta(x_0)(x - x_0) = \alpha(x) - \alpha(x_0) - T(x - x_0)$. Since the left side is tangent at $x = x_0$ for x ranging over E, it is also tangent for x ranging over F. Hence, the right side is also tangent at $x = x_0$, x ranging over F. Thus, α is differentiable at $x = x_0$, and $D\alpha(x_0) = T$. Similarly, if β is C^p, then α is also C^p.

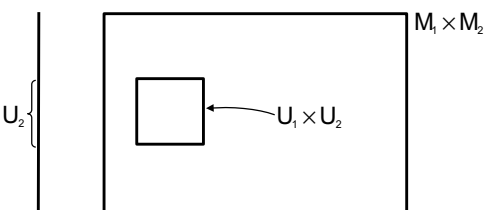
Thus, we have obtained a collection of $F-$charts on the set N, satisfying the first, second, and forth condition for a manifold. By the Lemma, therefore, we have a manifold N based on F. The manifold so obtained is called a *submanifold* of M. [For some reason that I do not understand, this term is normally reserved for the case when F splits in E.]

Example. Let M be a a manifold based on Banach space E, and let O be an open subset of M. We claim that this O is a sub manifold of M, i.e., that O satisfies the condition on N at the top of the previous page (with $F = E$). Indeed, let p be any point of O, and let U, ψ be any admissible chart with p in U. Then, setting $U' = U \cap O$ and letting ψ' be the restriction of ψ to U', we obtain an admissible chart U', ψ'. But, for this chart, we have $\psi'[U' \cap O] = \psi'[U'] \cap E$. Thus, we obtain a manifold O based on E.

Example. Regard the manifold M of the third example on page 25 as a subset of the manifold E (consisting of $(r_1, r - 2, \ldots)$ the sum of whose squares converges). Then M is a sub manifold.

We turn, finally, to the second method for constructing manifolds from manifolds. Let M_1 and M_2 be C^p manifolds, based on Banach spaces E_1 and E_2, respectively. We obtain a new manifold. Let $M = M_1 \times M_2$, Cartesian product of sets (so an element of M is a pair, (m_1, m_2), with m_1 in M_1 and m_2 in M_2). Next, let $E = E_1 \times E_2$, product of Banach spaces. we now introduce some $E-$charts on this set M. Let U_1, ψ_1 and U_2, ψ_2 be admissible charts on M_1 and M_2, respectively. Set $U = U_1 \times U_2$, a subset of M (so (m_1, m_2) in M is in U provided m_1 is in U_1 and m_2 is in U_2). Let ψ be the mapping from U to E with action $\psi(m_1, m_2) = (\psi_1(m_1), \psi_2(m_2))$, noting that the object on the right is indeed in E.

We now claim that this U, ψ is an $E-$chart on M. Indeed, ψ is one-to-one, for $\psi(m_1, m_2) = \psi(m'_1, m' - 2)$ implies $\psi_1(m_1) = \psi_1(m'_1)$ and $\psi_2(m_2) =$



$\psi_2(m'_2)$, whence, since ψ_1 and ψ_2 are one-to-one, we have $m_1 = m'_1$ and $m_2 = m'_2$, whence $(m_1, m'_2) = m_2, m'_2)$. Next, note that $\psi[U]$ is open in E, for $\psi[U]$ is the subset $\psi_1[U_1] \times \psi_2[U_2]$ of E, while each factor is open in its respective Banach space. Thus, U, ψ is indeed a chart on M.

We have so far a set M, together with some charts on this set. We claim, next, that this collection of charts satisfies the first, second, and fourth conditions for a manifold. For the second condition, note that, given (m_1, m_2) in M, we have, choosing U_1, ψ_1 with m_1 in U_1 and U_2, ψ_2 with m_2 in U_2, that (m_1, m_2) is in the corresponding U. Similarly, the fourth condition follows from that condition on M_1 and M_2. Thus, we have only to verify the first condition, i.e., to show that any two of our charts are compatible. LetU, ψ and U', ψ' be two such, coming from U_1, ψ_1 and U_2, ψ_2 and U'_1, ψ'_1 and U'_2, ψ'_2, respectively. Then, setting $V_1 = U_1 \cap U'_1$ and $V_2 = U_2 \cap U'_2$, we have that $U \cap U' = V_1 \times V_2$. Thus, $\psi^{-1}[U \cap U' = \psi_1^{-1}[V_1] \times \psi_2^{-1}[V_2]$, whence, since each factor on the right is open in its Banach space, $\psi^{-1}[U \cap U']$ is open in E, and similarly for $\psi'^{-1}[U \cap U']$. Next, consider the mapping $\psi' \cdot \psi^{-1}$ from $\psi[U]$ to E. It is action on $x_1, x_2)$ in E (so x_1 is a vector in E_1 and x_2 a vector in E_2) is $\psi' \cdot \psi^{-1}(x_1, x_2) = (\psi'_1 \cdot \psi_1^{-1}(x_1), \psi'_2 \cdot \psi_2^{-1}(x_2))$. But, since $\psi'_1 \cdot \psi_1^{-1}$ (from $\psi_1[U_1]$ to E_1) is C^p and $\psi'_2 \cdot \psi_2^{-1}$ is C^p, so is this $\psi' \cdot \psi^{-1}$. We conclude that, indeed, any two of our charts are C^p-compatible.

Thus, we so far have a set M, and a collection of $E-$charts on M satisfying the first, second, and fourth conditions for a manifold. By the Lemma, therefore, we have a C^p manifold M based on E. This manifold is called the *product* of M_1 and M_2, written $M_1 \times M_2$.

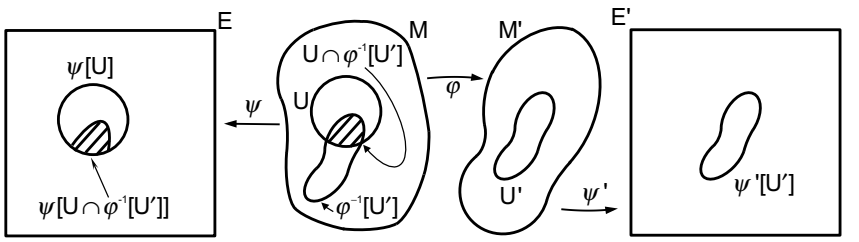
Example, Let E_1 and E_2 be Banach spaces. Then the product of the manifold E_1 (second example, page 25) with the manifold E_2 is the manifold $E_1 \times E_2$ (product of Banach spaces).

Example Let M_1 and M_2 be C^p manifolds, and let M be the C^p manifold $M_1 \times M_2$. Fix once and for all a point $\underline{m_2}$ of M_2. Denote by N the subset of M consisting of all elements of the form $(m_1, \underline{m_2})$ with m_1 in M_1. Then N is a submanifold of M (where the corresponding subspace F of E is just the subspace E_1 of E). Clearly, the resulting manifold N (based on E_1) is just a copy of the original manifold M_1. In this sense, then, $M_1 \times M_2$ is "sliced by submanifolds which are copies of M_1, and also by copies of M_2", just as one might expect of a product.

A characteristic feature of these two constructions should be noted. A manifold "looks locally like a certain Banach space". In each case, one simply takes a construction applicable to Banach spaces ("taking a subspace" and "taking a product", respectively) and performs essentially that same construction for manifolds, using the charts to "pull over" the construction from Banach spaces to the manifolds. This general theme persists throughout the subject; Things done to or on Banach spaces, "localized and pulled over via charts", yield things done to or on manifolds.

10. Mappings of Manifolds

Each kind of mathematical object normally comes equipped with the notion of a "structure-preserving mapping" between two such objects. For vector spaces, the mappings are linear mappings; for groups, homomorphisms, for topological spaces, continuous mappings; for Banach spaces, bounded linear mappings. Manifolds are a "kind of mathematical object". We now introduce the corresponding mappings.



Let M and M' be C^p manifolds, based on Banach spaces E and E', respectively. Let φ be a mapping from set M to set M'. This φ is said to be a C^p *mapping* (of manifolds) provided the following condition is satisfied: Given any admissible charts U, ψ and U', ψ' on M and M', respectively, i) the subset $\psi[U \cap \varphi^{-1}[U']]$ of E is open, and ii) the mapping $\psi' \cdot \varphi \cdot \psi^{-1}$ from this $\psi[U \cap \varphi^{-1}[U']]$ to E' is C^p. We first check that these subsets and mappings make sense. First, $\varphi^{-1}[U']$ is some subset of M. Hence, $U \cap \varphi^{-1}[U']$ is some possibly smaller subset of M, and indeed is subset of U. Hence, $\psi[U \cap \varphi^{-1}[U']]$ is well-defined, and is a subset of E. For x in this subset, $\psi^{-1}(x)$ is in $U \cap \varphi^{-1}[U']$, and so, in particular, $\psi^{-1}(x)$ is in $\varphi^{-1}[U']$. Hence, $\varphi(\psi^{-1}(x))$ makes sense, and is an element of M' – indeed, is actually an element of the subset U' of M'. Hence, $\psi'(\varphi(\psi^{-1}(x)))$ makes sense and is an element of E'. Thus, everything makes sense. In condition i), let us fix U', ψ', and vary U, ψ. Then condition i) requires precisely that $\varphi^{-1}[U']$ be an open subset of M. In topological terms, a C^p mapping must be a continuous mapping of topological spaces. In the case – $p = 0$, this is the only condition, for condition ii) then follows (since compositions of continuous mappings are continuous) from condition i). For other p, however, we require still more: Roughly speaking, we require that, "if the mapping φ is pulled back via the charts to a mapping of Banach spaces (so we know what C^p means), then the result is a C^p mapping between those Banach spaces". Thus, this definition can be viewed as another instance of our general program: We know what "C^p" mapping" means for Banach spaces, and we carry over that notion to manifold via charts.

Example. Let M be a C^p manifold, and N a submanifold, so N is also a C^p manifold in its own right. Let φ be the mapping from set N to set M with $\varphi(n) = n$, i.e., which "inserts N into M". Then this φ is a C^p mapping of manifolds.

Example. Let M_1 and M_2 be C^p manifolds, so $M_1 \times M_2$ is also a C^p manifold. Let φ be the mapping from $M_1 \times M_2$ to M_1 with action $\varphi(m_1, m_2) = m_1$ ("projection onto the first factor"). Then φ is a C^p mapping of manifolds.

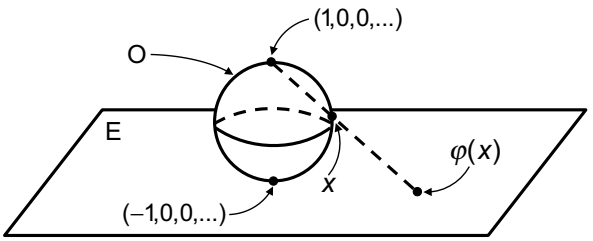
Example. Let M be a C^p manifold, and consider the C^p manifold $\mathbb{R}$ (a Banach space, and hence a C^∞ manifold, and hence a C^p manifold). A C^p mapping γ from R to M is called a *curve* in M.

Example Let M be a C^p manifold, and let U, ψ be an admissible chart on M. Set $K = \psi[U]$, so K is an open subset of Banach space E. We may regard K as C^p manifold (since E, as a Banach space, is a manifold, and K is an open subset of E: first example, page 27). Then ψ is a mapping from manifold K to manifold M. This is a C^p mapping.

Let M and M' be C^p manifolds, and let φ be a C^p mapping from M to M'. If φ happens to have an inverse, i.e., if there is a C^p mapping λ from M' to M such that $\lambda \cdot \varphi$ is the identity mapping on M and $\varphi \cdot \lambda$ is the identity mapping on M', then φ is

called a *diffeomorphism* (or C^p diffeomorphism) from M to M'. Then necessarily λ is a diffeomorphism from M' to M. A diffeomorphism from M to M' "makes M and M' identical as manifolds" (i.e., is analogous to an isomorphism of groups or of Banach spaces, a homeomorphism of topological spaces, etc.). If there exists a diffeomorphism from M to M' then M and M' are said to be *diffeomorphic* (or C^p diffeomorphic).

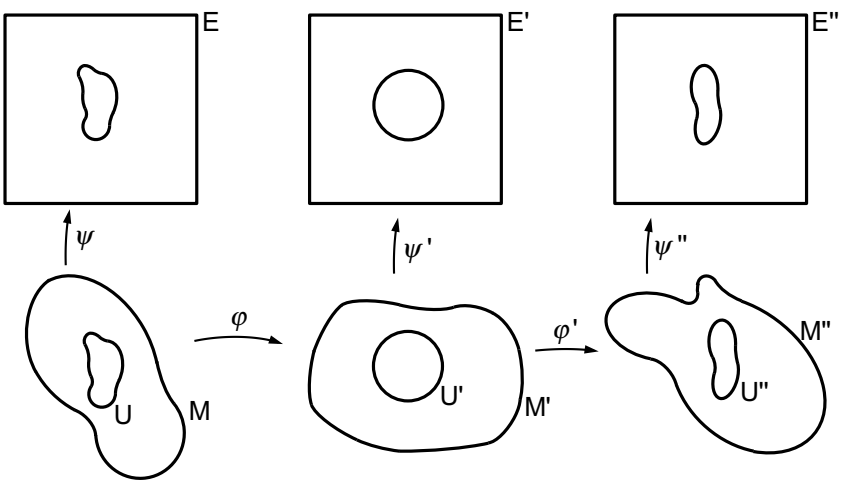
Example. Consider the manifold M of the third example on page 25. Let O be the subset of M consisting of all points thereof except $(1, 0, 0, \ldots)$. Then O, as an open subset of M, is also a manifold. The Banach space E of that example is, as a Banach space, also a manifold. Let φ be the mapping from O to E which sends $(r_1, e_2, \ldots)$ in) to $(1/(1 - r_1))(r_2, r_3, \ldots)$ in E. This is a diffeomorphism of manifolds. [For example, a two-dimensional sphere, minus a point, is diffeomorphic with the plane.]



The composition of two C^p mappings of Banach space is C^p. As one expects, this carries over directly to manifolds.

Theorem. Let M, M', and M'' be C^p manifolds, and let φ be a C^p mapping from M to M', and φ' a C^p mapping from M' to M''. Then the mapping $\varphi' \cdot \varphi$ from M to M'' is also a C^p mapping of manifolds.

Proof: Let E, E', and E'' be the Banach spaces on which these manifolds are based. Choose charts, U, ψ, U', ψ', and U'', ψ''. Then, since $\psi'[U' \cap \varphi'^{-1}[U'']]$ is open in E', for every $U', \psi', \psi[U \cap (\varphi' \cdot \varphi)^{-1}[U'']]$ is open in E. Since the mapping of Banach spaces $\psi' \cdot \varphi \cdot \psi^{-1}$ and $\psi'' \cdot \varphi' \cdot \psi'^{-1}$ are C^p, so is their composition, $\psi'' \cdot \varphi' \cdot \varphi \cdot \psi^{-1}$.



Example The composition of two diffeomorphism is a diffeomorphism, for, if λ and λ' are the inverses of φ and φ', respectively, than $\lambda \cdot \lambda'$ is the inverse of $\varphi' \cdot \varphi$, and C^p-ness is preserved under composition. Hence, "is diffeomorphic with" is an equivalence relation on manifolds.

Example. Given a curve γ on M (so γ is a C^p mapping from $\mathbb{R}$ to M), and a C^p mapping φ from M to M', then $\varphi \cdot \gamma$, as a C^p mapping from R to M', is a curve on M'.

11. Scalar Fields

We have now completed, in the last two sections, our discussion of what a manifold is and is like: We have the notion of a manifold, and of a structure-preserving mapping between manifolds. It turns out, however, that manifolds, viewed as simply objects by themselves, are not all that interesting. What is perhaps more interesting is the various objects which live naturally in the environment of a manifold. Among these objects are what will be called tensor fields. We now begin, therefore, a program of defining and finding the properties of these tensor fields. Our approach will be to first treat scalar fields and vector fields (by far the two most important examples). These two examples out of the way, we shall then look at the entire situation regarding tensor fields from a more systematic viewpoint. We begin, then, with the scalar fields.

Fix a C^p manifold M, based on Banach space E. A *scalar field* (or C^p scalar field) on M is a C^p mapping f from M to R (where R is here regarded as a C^p manifold). That is, a scalar field on M is just a real-valued function on M, a function which happens to be smooth in a certain sense.

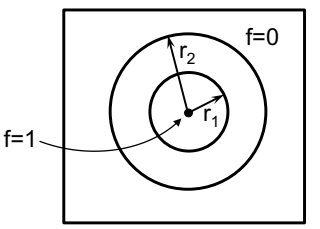
Example. Regard Banach space E as a C^p manifold. Then a bounded linear mapping from E to $\mathbb{R}$ is a scalar field on E.

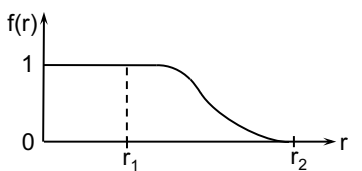
Example. Let f be a C^p scalar field on M, and γ a C^p curve. Then $f \cdot \gamma$ is one real function of one real variable (representing "evaluation of the scalar field along the curve"). As composition of C^p mappings of manifolds, this function is also C^p.

Fix C^p manifold M. Let f and f' be two C^p scalar fields on M, and consider the function $f + f''$ of M with action $(f + f'')(p) = f(p) + f'(p)$. Then this function is also a C^p scalar field (since the sum of two C^p functions on a Banach space is a C^p function). Furthermore, for a a real number, the function af with action $(af)(p) = a\,f(p)$ is a C^p scalar field. That is to say, the set of all C^p scalar fields on M has the structure of a real vector space. Finally, for f and f' C^p scalar fields , the function $f\,f'$ with action $(f\,f')(p) = f(p)\,f'(p)$ is also C^p (since true on Banach spaces). Thus, we have available a "multiplication" within our vector space. The collection of all C^p scalar fields on M thus becomes a ring (and, indeed, a commutative, associative algebra with unit).

Next, we consider the behavior of scalar fields under mappings of manifolds. Let M and M' be C^p manifolds, and let φ be a C^p mapping from M to M'. Then, given any C^p scalar field f on M', we have that $f \cdot \varphi$ is a real-valued function on M. But, as the composition of C^p mappings, this function is also C^p, and hence is a C^p scalar field on M. Thus, we can "pull back" scalar fields from M' to M under mappings from M to M'. Clearly, the "pull-back" of the sum of two fields is the sum of their pull-backs, the pull-back of a numerical multiple is that multiple of the pull-back, and the pull-back of the product is the product of the pull-backs. In other words, φ, a structure-preserving mapping from M to M', defines a structure-preserving mapping from scalar fields on M' to those on M. [In algebraic language, this last mapping is a homomorphism of rings.]

Finally, we give some examples of some interesting – and perhaps a bit surprising – difference between the properties of scalar fields in the finite – and infinite-dimensional cases. Consider first the following statement: for E a Banach space, and r_1 and r_2 positive numbers with r_1, r_2 there is a C^∞ real-valued function f on E with $f(x) = 1$ whenever $|x| \leq r_1$ and $f(x) = 0$ whenever $|x| \geq r_2$. We first note that this statement is true if E is finite-dimensional: Indeed, for $(r_1, \ldots, r_n)$ in R^n, set

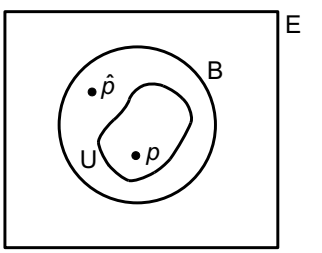




$r = [(r_1)^2 + \ldots + (r_n)^2]^{1/2}$, and let f be a C^∞ function of r which has value one for $r \leq r_1$ and zero for $r \geq r_2$. In fact, this statement in the finite-dimensional case represents an important and frequently used property of R^n: It allows one to "localize arguments" by choosing functions which do what one wants in some small region, but which, since they are zero farther away, do not do anything very nasty outside of that small region. For example, one can easily show from our fact that a finite-dimensional manifold admits a "reasonable number" of C^∞ scalar fields, in, e.g., the following sense: Given distinct points p and $p!$ of finite-dimensional M, and any numbers a and a', there exists a C^∞ scalar field f on M with $f(p) = a$ and $f(p') = a'$.

Is the statement of the previous paragraph true in general, i.e., is it also true in the infinite-dimensional case? It turns out that the answer is no.

Example. Let E be the Banach space of all sequences of reals the sum of whose absolute values converges, with norm this sum. Fix positive r, and let f be a C^1 scalar field on the ball B of radius r centered at 0. Also, let U be an open subset of E whose closure, $\overline{U}$ is in B. Then, given any point p of U and any positive ϵ, there is a point $\hat{p}$ of B, not in U, such that $|f(\hat{p}) - f(p)| \leq \epsilon$.

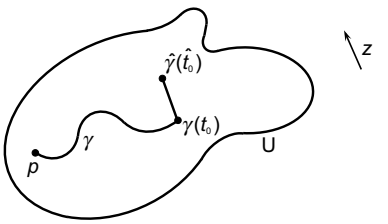


Proof: Nothing is lost be setting $r = 1$ and $f(p) = 0$. Fix, once and for all, positive number ϵ. Denote by Γ the collection of all pairs (t_0, γ), where t_0 is a non-negative number and γ is a continuous mapping from the closed interval $[0, t_0]$ to $\overline{U}$, satisfying the following conditions: i) $\gamma(0) = \beta$, and $|\gamma(t_0) - \gamma(0)| \geq t_0/2$, ii) for any t and t' in $[0, t_0]$, $|\gamma(t) - \gamma(t')| \leq |t - t'|$, and iii) $f(\gamma(t_\iota)) \leq \epsilon t_0$. [Think of the curve as a "moving point in $\overline{U}$ whose location at time t is $\gamma(t)$". The first condition then requires that "the point begin at p, and manage to achieve distance at least $t_0/2$ from p by time $t-0$"; the second that "the speed of the moving point not exceed one"; the third that "the function f not be too large where the curve ends.] We now partially order this set Γ by the relation "is an extension of": $(t_0, \gamma) \leq (t'_0, \gamma')$ provided $t_0 \leq t'_0$ and $\gamma = \gamma'$ on $[0, t_0]$. Note that, by condition i) and the fact that $\overline{U}$ is a subset of a ball of radius one, we have $t_0 \leq 4$ for every (t_0, γ) in Γ.

We next claim that this partially ordered set Γ satisfies the condition for Zorn's Lemma. Indeed, given any totally ordered subset of Γ, we certainly obtain, by "stringing these extensions together", a t_0 (the lub of the $y'_0 s$ of the totally ordered subset), and a continuous mapping γ from $[0, t_0]$ to $\overline{U}$ satisfying the first part of condition i) and condition ii). But, by condition ii), the sequence $\gamma(t_0/2), \gamma(3t_0/4), \gamma(7t_0/8), \ldots$ in $\overline{U}$ is Cauchy, whence, since $\overline{U}$ is closed, this sequence converges to some element of $\overline{U}$. Denote this element $\gamma(t_0)$, so we now have γ a mapping from $[0, t_0]$ to $\overline{U}$. But this (t_0, γ) is in Γ, the three conditions following from continuity. Thus, we have obtained our upper bound, and so verified the condition for Zorn's Lemma.

By Zorn's lemma, there is a maximal element, (t_0, γ), of Γ. If $\gamma(t_0)$ (necessarily in $\overline{U}$ were not in U, then we would be done (for, setting $\hat{p} = \gamma(t_0)$, we would have $|f(\hat{p})| \leq \epsilon t_0 \leq 4\epsilon$, where the first inequality is from condition iii) and the second from $t_0 \leq 4$). Thus, we have only to show that assumption that $x_0 = \gamma(t_0)$ is in U leads to a contradiction. Make this assumption, and first note that, since f is C^1 at x_0, there is a positive δ such that, whenever $|z| \leq \delta, |f(x_0 + z) - f(x_0) - Df(x_0)(z)| \leq \epsilon\delta$. We next claim that there is a vector z in E such that $|z| = \delta, |x_0 - p + z| \geq |x_0 - p| + \delta/2$, and $Df(x_0)(z) = 0$. [Proof: Since $x_0 - p$ is in E, it is represented by a sequence, $(r_1, r_2, \ldots)$, of reals the sum of whose absolute values converges. Choose two entries of this sequence, say r_2 and r_5, with $|r_2| \leq \delta/4$ and $|r_5| \leq \delta/4$. Consider the subspace of E consisting of elements all of whose entries, except possibly the second and fifth, are zero. Since $Df(x_0)$ is a linear mapping from this subspace to the reals, and since this subspace is two-dimensional,

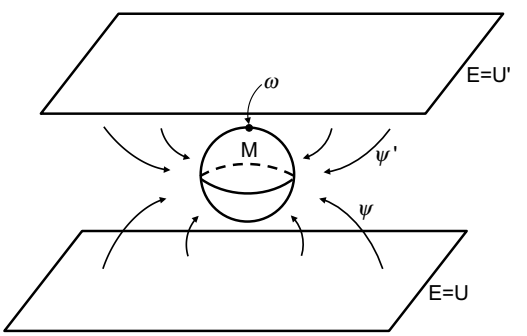
there is some element, z, of this subspace with $|z| = \delta$ and $Df(x_0)(z) = 0$. But now, since z is of the form $(0, u, 0, 0, v, 0, \ldots)$, and since $x_0 - p = (r_1, e_2, \ldots)$ has $|r_2| \leq \delta/4$, we have also $|x_0 - p + z| \geq |x_0 - p| + \delta/2$. This z, then, is what we wanted.]



Now set $\hat{t_0} = t_0 + \delta$, and let $\hat{\gamma}$ be the mapping from $[0, \hat{t_0}]$ with $\hat{\gamma}(t) = \gamma(t)$ for t in $[0, t_0]$, and $\hat{\gamma}(t) = \gamma(t_0) + (t - t_0)/\delta\ z$ for $t \geq t_0$. We show that this $(\hat{t_0}, \hat{\gamma})$ is also in Γ, violating maximality of (t_0, γ), and thus giving us our desired contradiction. Condition i): Clearly, $\hat{\gamma}(0) = p$. Also, $|\hat{\gamma}(\hat{t_0}) - \hat{\gamma}(0)| = |\gamma(t_0) + z - \gamma(0)| = |x_0 - p + z| \geq |x_0 - p| + \delta/2 \geq 1/2(t_0 + \delta) = 1/2\hat{t_0}$, where we used a property of z above in the third step, and condition i) on (t_0, γ) in the fourth. Condition ii):

For, e.g., t in $[t_0, \hat{t_0}]$ and t' in $[0, t_0]$, we have $|\hat{\gamma}(t) - \hat{\gamma}(t')| = |\gamma(t_0) + (t - t_0)/\delta z - \gamma(t')| \leq |\gamma(t_0) - \gamma(t')| + (t - t_0)/\delta\delta \leq (t_0 - t') + (t - t_0) = |t - t'|$, where we used $|z| = \delta$ in the second step, and condition ii) on (t_0, γ) in the third. Condition iii): Since $Df(x_0)(z) = 0$ and $|z| = \delta$, we have, by the paragraph above, that $|f(x_0 + z) - f(x_0)| \leq \epsilon\delta$. Hence, $|f(\hat{\gamma}(\hat{t_0}))| = |f(x_0 + z)| \leq |f(x_0)| + |f(x_0 + z) - f(x_0)| \leq \epsilon t_0 + \epsilon\delta = \epsilon\hat{t_0}$. This completes the proof of the result claimed in our example.

The crucial step in the proof above is finding the "direction z, motion along which gets one significantly further from p, but along which the function f does not increase too much". The statement itself asserts that "the C^1 function f tastes outside of U any value that it assumes inside of U". This behavior of C^1 functions in infinite dimensions is rather like the well-known behavior of harmonic functions in finite dimensions. One concludes, then, that C^1 functions on certain infinite-dimensional spaces are similar in their behavior to harmonic functions on finite-dimensional spaces. Indeed, we can push this analogy still further. The only harmonic functions on a finite-dimensional sphere are the constants. In infinite dimensions, we have the following.



Example. Let E be the Banach space of the example above. Let $\tilde{E}$ be a copy of the set E (with φ a one-to-one, onto mapping from $\tilde{E}$ to E), and let M be the union of $\tilde{E}$ with one additional point ω We introduce two $E-$charts on the set M. For one, set $U = \tilde{E}$, and $\psi = \varphi$. For the other, set U' the subset of M consisting of ω, together with those κ in $\tilde{E}$ with $\varphi(\kappa) =\neq 0$ and let $\psi'(\omega) = 0$ and $\psi'(\kappa)\varphi(\kappa)/|\varphi(\kappa)|^2$, for κ in $\tilde{E}$. These two charts make M a C manifold based on E. [This M is "like a sphere", in, e.g., the sense that the same construction in the finite-dimensional case yields a sphere.]

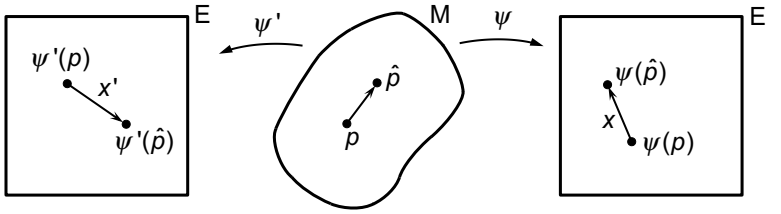
We now claim that the only C^1 scalar fields f on this manifold M are those which are constant. Indeed, let f be such a field, and let $f(\omega) = a$. Then $f \cdot \psi'^{-1}$ is a C^1 function on E, with $f \cdot \psi'^{-1}(0) = a$. Fix positive ϵ. Then, by continuity, there is a positive δ such that whenever $|x| \leq \delta$, $|f \cdot \psi'^{-1}(x) - f \cdot \psi'^{-1}(0)| \geq \epsilon$. Now, $f \cdot \psi^{-1}$ is a C^1 function on E. By what we have just shown, whenever $|x| \leq 1/\delta, |f \cdot \psi^{-1}(x) - a| \leq \epsilon$. But, by the result of the previous example, it follows that $|f \cdot \psi^{-1}(x) - a| \leq \epsilon$ for every x in E. Hence, $|f(p) - a| \leq \epsilon$ for every p in M. Since ϵ is arbitrary, we must have $f(p) = a$ for every p in M.

Thus, the manifold of this example, at least, has no interesting C^1 scalar field whatever.

12. Vector Fields

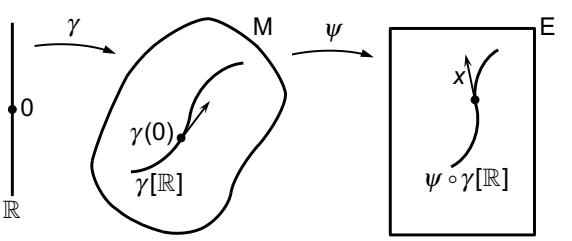
We have now completed our discussion of the first example of a field on a manifold: scalar fields. We turn now to the second example.

Let M be a C^p manifold ($p \geq 1$) based on Banach space E. Fix a point p of M. Consider now the collection of all pairs, $(x; U\psi)$, where x is a vector in E and U, ψ is an admissible chart on M with p in U. Given two such, write $(x; U, \psi) \approx (x'; U', \psi')$ if $D(\psi' \cdot \psi^{-1})(\psi(p)(x) = x'$. We now claim that this $\approx$ is an equivalence relation. [Proof: Clear, $(x; U, \psi)$, $\approx (x; U, \psi)$. Next, let $(x; U, \psi) \approx (x'; U', \psi')$. We have that $(\psi' t \psi^{-1}) \cdot (\psi \cdot \psi'^{-1})$ is the identity mapping. Taking the derivative of this equation, using the chain rule, we have $D(\psi' \cdot \psi^{-1})(\psi(p)) D(\psi \cdot \psi'^{-1})(\psi'(p)) = I$, the identity on E. Hence, since $D(\psi' \cdot \psi^{-1})(\psi(p))(x) = x'$ we have $D(\psi \cdot \psi'^{-1})(\psi'(p))(x') = x$, i.e., $(x'; U', \psi') \approx (x; U, \psi)$. Finally, let $(x; U, \psi) \approx (x'; U', \psi')$ and $(x'; U', \psi') \approx (x''; U'', \psi'')$. Taking the derivative of $(\psi'' \cdot \psi'^{-1}) \cdot (\psi' \cdot \psi^{-1}) = \psi'' \cdot \psi^{-1}$, we have that $D(\psi'' \cdot \psi'^{-1})(\psi'(p)) D(\psi' \cdot \psi^{-1})(\psi(p)) = D(\psi'' \cdot \psi^{-1}(\psi(p))$. Applying each side to x, we have $D(\psi'' \cdot \psi^{-1})(\psi(p))(x) = D(\psi'' \cdot \psi'^{-1})(\psi'(p))(D(\psi' \cdot \psi^{-1}(\psi(p))(x)) = D(\psi'' \cdot \psi'^{-1})(\psi'(p))(x') = x''$. hence, $(x; U, \psi) \approx (x''; U'', \psi'')$.] An equivalence class is called a *tangent vector* (to M, at p), or just vector in M.



Intuitively, a tangent vector at p represents "an infinitesimal displacement from p, in M". We support this intuitive picture, and thus motivate the definition above, by the following "first order calculation". Fix point $\hat{p}$ in M "near p". Let us represent the relationship between these points in terms of chart U, ψ, with p in U. This representation can be accomplished by considering the vector $x = \psi(\hat{p}) - \psi(p)$ in E. A different chart gives a different representation, namely, U', ψ' gives $x' = \psi'(\hat{p}) - \psi'(p)$. To compare these two "representations of the same displacement in M", we note that $x' = \psi'(\hat{p}) - \psi'(p) = (\psi' \cdot \psi^{-1})(\psi(\hat{p})) - (\psi' \cdot \psi^{-1})(\psi(p)) \approx D(\psi' \cdot \psi^{-1})(\psi(p))(\psi(\hat{p}) - \psi(p)) = D(\psi' \cdot \psi^{-1})(\psi(p))(x)$, where "$\approx$" means "ignoring a mapping tangent at $\psi(p)$". But this formula is precisely the equivalence relation above. The "term ignored becomes negligible as the displacement between $\hat{p}$ and p goes to zero". Thus, we interpret our equivalence relation as requiring that "various representation in terms of charts all correspond to the same infinitesimal displacement in M".

Further support for our intuitive picture comes from the following.



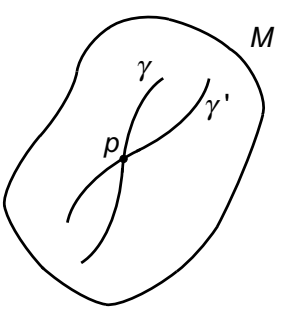
Example. Let M be a C^p manifold ($p \geq 1$), and let γ be a curve in M, i.e., a C^p mapping from $\mathbb{R}$ to M. Let p be the point $\gamma(0)$ of M. We obtain a tangent vector at p. Given any chart U, ψ in M with p in U, we have that $\psi \cdot \gamma$ is a mapping from $\mathbb{R}$ to E, whence its derivative at $0, D(\psi \cdot \gamma)(0)$ is a linear mapping from $\mathbb{R}$ to E, whence $x = D(\psi \cdot \gamma)(0)(1)$ is a vector in E. Given another such chart, we similarly set $x' = D(\psi' \cdot \gamma)(0)(1)$. But, taking the derivative of $\psi' \cdot \gamma = (\psi' \cdot \psi^{-1}) \cdot (\psi \cdot \gamma)$ at 0, we have $D(\psi' \cdot \gamma)(0) = D(\psi' \cdot \psi^{-1})(\psi \cdot \gamma(0))\ D(\psi \cdot \gamma)(0)$. applying this to the number "1" in $\mathbb{R}$, we see that $x' = D(\psi' \cdot \psi^{-1}(\psi(\gamma(0)))(x)$. That is to say, we have that $(x; U, \psi) \approx (x'; U', \psi')$.

We thus obtain in this way an equivalence class as on the previous page. This tangent vector is called the *tangent* to the curve γ at $\gamma(0)$.

This last example gives what we might expect: Since a curve "moves in M", it should define, at each point of the curve, a tangent vector giving the "direction of motion in M". Example. Let M be a C^p manifold ($p \geq 1$), p a point of M, ξ a tangent vector at p, and f a C^p scalar field on M. Let $(x; U, \psi)$ be any representative of the equivalence class ξ, and consider the number $a = D(f \cdot \psi^{-1})(\psi(p))(x)$. Given any other $(x'; U', \psi')$ in this equivalence class, we similarly define a'. But, taking the derivative of the identity $(f \cdot \psi'^{-1})(\psi' \cdot \psi^{-1}) = (f \cdot \psi^{-1})$, and using the fact that $(x; U, \psi) \approx (x'; U', \psi')$, we have that $a = a'$. Thus, we obtain, from vector ξ and scalar field f, a unique number, independent of choice of chart used to define this number. This number is called the *directional derivative* of f in the direction ξ. Intuitively, this number is " the rate of change of f under the infinitesimal displacement in M defined by ξ".

The set of all tangent vectors at p is called the *tangent space* (to M) at p. Our next goal is to determine what structure there is on this set. To this end, we first note the following fact: If ξ is any tangent vector at p, and U', ψ' is ant chart with p in U', then there is one and only one vector x' in E with $(x'; U', \psi')$ in the equivalence class ξ. Indeed, for $(x; U, \psi)$ any representative of ξ, the required unique x' is that given by $x' = D(\psi' \cdot \psi^{-1})(\psi(p))(x)$. Thus, given a chart U, ψ with p in U, we obtain a one-to-one, onto mapping α from the tangent space at p to the Banach space E (namely, $\alpha(\xi)$ is that x in E such that $(x; U, \psi)$ is in the equivalence class ξ). This correspondence α thus induces on the tangent space all the structure of the Banach space E (i.e., a vector-space structure and norm structure). We are of course interested only in that part of all this structure on the tangent space which is independent of the choice of chart. So, let U', ψ' be another. Then we obtain immediately (since, for any $\xi, (\alpha(\xi); U, \psi') \approx (\alpha'(\xi) : U'\psi')$) that $\alpha' = D(\psi' \cdot \psi^{-1})(\psi(p)) \cdot \alpha$. Clearly, then, the vector-space structure on the tangent space is chart-independent (i.e., if $\alpha(\tau) = \alpha(\xi) + \alpha(\eta)$, then $\alpha'(\tau) = \alpha'(\xi) + a\alpha'(\eta)$). The norm structure, however, does depend on the chart, for $|\alpha'(\xi)| = |D(\psi' \cdot \psi^{-1})(\psi(p))(\alpha(\xi))|$, which is not in general equal to $|\alpha(\xi)|$. We have, however, that $|\alpha'(\xi)| \leq |D(\psi' \cdot \psi^{-1})(\psi(p))|\,|\alpha(\xi)|$, and similarly, reversing the roles of α and α'. That is to say, any two norms obtained on the tangent space, from two charts, are equivalent to each other: It is only the actual numerical values of the norms which are chart-dependent. This structure on the tangent space – a vector space, together with a collection of norms theorem which make the vector space a Banach space and which are all equivalent to each other – is called a Banachable space. We can, in the tangent space, speak of sums of tangent vectors, numerical multiples of tangent vectors, limits of a sequence of tangent vectors, Cauchy sequences of tangent vectors (all notions which depend only on the vector-space structure or on the norm up to equivalence): We cannot, for example, speak of the norm of a tangent vector (as a real number).

In the finite-dimensional case, there are a number of equivalent definitions of tangent vectors. It is of some interest to see which of these definitions agree with ours above in general (i.e., also in infinite dimensions). We consider two.



Let M be a C^p manifold ($p \geq 1$), and fix a point p of M. Given any C^p curve γ on M with $\gamma(0) = p$, and any C^p scalar field f on M, denote by $s(\gamma, f)$ the number $D(f \cdot \gamma)(0)(1)$ (l.e., the ordinary derivative, at zero, of the real function $f \cdot \gamma$ of one real variable). [In our language, this is the directional derivative of f in the direction of the tangent to γ at zero.] One then calls two curves, γ and γ', equivalent if $s(\gamma, f) = s(\gamma', f)$ for every f (i.e., if, in intuitive terms, the two curves are "tangent to each other at p"). Tangent vectors are then defined as the equivalence classes. This definition is not, in general, the same as ours, but for a minor technical reason: As we saw in the previous section, there are manifolds which admit only the constant C^p scalar fields, and for these,

for example, we would obtain only one equivalence class (since we would have $s(\gamma, f) = 0$ always). We can avoid this difficulty, however, by considering instead C^p scalar fields defined on a sufficiently small open sub manifold of M containing p. [Choosing, e.g., this open subset to be the "U" of a chart, we will obtain enough C^p scalar fields.] With this one modification, the two definitions will coincide. [Any tangent vector in this sense defines one in our sense, since equivalent curves have the same tangent. Conversely, give a tangent vector in our sense, one constructs a curve whose tangent at p is that tangent vector (using a chart), and thus obtains an equivalence class of curves and hence a tangent vector in this sense.]

For the second definition, again let M be a C^p manifold ($p \geq 1$), and let p be a point of M. Denote by $\mathfrak{J}$ the collection of all C^p scalar fields on M, so $\mathfrak{J}$ is a vector space with products. A derivation on $\mathfrak{J}$ is a mapping δ from $\mathfrak{J}$ to the reals, satisfying the following conditions: i) For f and f' in $\mathfrak{J}$, $\delta(f + f') = \delta(f) + \delta(f')$, ii) For f a constant function, $\delta(f) = 0$, and iii) For f and f' in $\mathfrak{J}\delta(ff') = f(p)\delta(f') + f'(p)\delta(f)$. As an example of a derivation, we have the following: Fix a tangent vector ξ to M at p, and denote by δ the mapping from $\mathfrak{J}$ to $\mathbb{R}$ with the following action: For f in $\mathfrak{J}, \delta(f)$ is the number given by the directional derivative of f in the ξ direction. Then this δ is a derivation (the three properties above being just properties of the directional derivative). For the second definition, tangent vectors are defined as derivations on $\mathfrak{J}$. This definition also does not agree with ours, for the same technical reason as above: Not enough scalar fields in general. Again, we can avoid this difficulty by choosing for $\mathfrak{J}$ the scalar fields in some small open set containing p. This having been done, does the present coincide with our original definition? It turns out that the answer is still no.

Example. Let the manifold M be the Banach space E itself (regarded as a manifold), and let the point p be the zero vector in E. Then, since M comes equipped naturally with a chart, the tangent space at p can be identified with E. Given any scalar field f on M, denote by τ_f the element of $\mathcal{L}(E; \mathbb{R})$ which sends x in E to the real number the directional derivative of f in the x direction (regarding x as a tangent vector at 0. Then, clearly, $\tau_{f+f'} = \tau_f + \tau_{f''}$ for f a constant scalar field, $\tau_f = 0$, and $\tau_{ff'} = f(0)\,\tau_{f'} + f'(0)\tau_f$.

Now let α be any element of $\mathcal{L}(\mathcal{L}(E; \mathbb{R}); \mathbb{R})$. Then for each f $\alpha(\tau_f)$ is just a real number. Hence, the mapping δ from $\mathfrak{J}$ to $\mathbb{R}$ with $\delta(f) = \alpha(\tau_f)$ is a derivative on $\mathfrak{J}$. Thus, so far we have a construction which yields, from an element α of $\mathcal{L}(\mathcal{L}(E; \mathbb{R}); \mathbb{R})$, a derivation. In particular, every element x of E determines a certain α_x in $\mathcal{L}(\mathcal{L}(E; \mathbb{R}) : R)$, namely, with the following action: For τ in $\mathcal{L}(E; \mathbb{R})$ let $\alpha_x(\tau) = \tau(x)$. Thus, since an element of E yields an element of $\mathcal{L}(\mathcal{L}(E; \mathbb{R}); \mathbb{R})$, and since an element of $\mathcal{L}(\mathcal{L}(E; \mathbb{R}); \mathbb{R})$ yields a derivation, each element of E yields a derivation. This, of course, is just the example we remarked on in the second paragraph of this page.

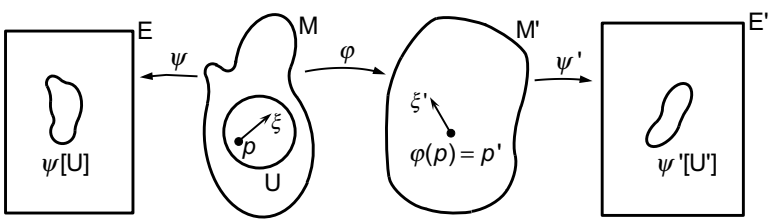
Is every α in $\mathcal{L}(\mathcal{L}(E; \mathbb{R}) : \mathbb{R})$ a α_x for some x in E? If not, we shall have an example of a derivation (namely, that which comes from this α) which arises from no tangent vector. We now find, for a specific choice of E, such an α.

Let E be the Banach space of all sequences of real numbers which converge to zero, with norm the lub of the absolute values of the entries. We now claim that $\mathcal{L}(E; \mathbb{R})$ is precisely the Banach space F of all sequences of reals the sum of the absolute values of whose entries converges (with norm this sum). Indeed, for $y = (s_1, s_2, \ldots)$ in F, let τ_y be the element of $\mathcal{L}(E; \mathbb{R})$ which sends $x = (r_1, r_2, \ldots)$ in E to the number $\tau_y(x) = s_1 r_1 + s_2 r_2 + \ldots$ (noting that the sum of the right converges, since the r_i converge to zero, and the sum of the absolute values of the s_i is finite). Conversely, let τ be an element of $\mathcal{L}(E; \mathbb{R})$. Set $s_1 = \tau(1, 0, \ldots)$, $s_2 = \tau(0, 1, 0, \ldots)$, etc. Then, for any $x = (r_1, r_2, \ldots)$ in E, we must have $\tau(x) = s_1 r_1 + s_2 r_2 + \ldots)$ (for the sequence in E whose n^{th} element is $(r_1, r_2, \ldots, r_n, 0, 0, \ldots)$ converges to x, whence τ of each element of this sequence must converge to $\tau(x)$. But τ of the n^{th} element above is $s_1 r_1 + \ldots + s_n r_n$). Furthermore, this $s_1, s_2, \ldots$ must be such that the sum of their absolute values converges. [Proof: Suppose not. Then, say , we have $|s_1| + \ldots + |s_7| \geq 1, |s_8| + \ldots + |s_1 3| \geq 1$, etc. Let $r_i, i = 1, \ldots 7$ each have absolute value $1/2$, and let the sign of r_i be the same as that of the

corresponding s_i; let $r_i, i = 8, \ldots 13$ each have absolute value $1/3$, each with the same sign as the corresponding s_i; etc. Then this $(r_1, r_2, \ldots)$ is in E, whereas $s_1 r_1 + s_2 r_2 + \ldots$ fails to converge: A contradiction.] Hence, $y = (s_1, s_2, \ldots)$ is an element of F. Clearly, we have $\tau = \tau_y$. We have shown, therefore, that every element of F defines an element of $\mathcal{L}(E; \mathbb{R})$, and conversely. That is, we have shown that $F = \mathcal{L}(E; \mathbb{R})$.

Now, finally, we are ready to choose our element α of $\mathcal{L}(\mathcal{L}(E; \mathbb{R}); |F|$. Let α have the following action on $\mathcal{L}(E; \mathbb{R}) = F$. For $y = (s_1, s_2, \ldots)$ in F, set $\alpha(y) = s_1 + s_2 + \ldots$ (noting that the sum converges, since the sum of the absolute values converges). For any $x = (r_1, r_2, \ldots)$ in E, the corresponding α_x has action $\alpha_x(y) = s_1 r_1 + s_2 r_2 + \ldots$. We now claim, finally, that there is no x in E such that $\alpha = \alpha_x$ (the only reasonable candidate being $x = (1, 1, 1, \ldots)$, which won't do, since this candidate is not in E). This completes our example.

The final topic, involving tangent vectors at a point, with which we shall deal is the question of their behaviour under mappings of manifolds. Let M and m' be C^p manifolds ($p \geq 1$), based on Banach spaces E and E', respectively. Let φ be a C^p mapping from M to M'. Fix a point p of M, set $p' = \varphi(p)$, and let ξ be a tangent vector to M at p. we obtain, using the mapping φ, a "corresponding" tangent vector at ξ' to M' at p'. [Intuitively, one expects such to be obtainable. Think of ξ as an "infinitesimal displacement at p in M". Then ϕ takes these "two nearby points in M to two nearby points in M', thus yielding a tangent vector at p' in M'".] The construction itself is this. Let U, ψ be a chart on M with p in U, and U', ψ' a chart on M' with p' in U'. Then $\psi' \cdot \varphi \cdot \psi^{-1}$ is a C^p mapping from an open subset of E to E'. Hence, $D(\psi' \cdot \varphi \cdot \psi^{-1})(\psi(p))$ is in $\mathcal{L}(E; E')$.



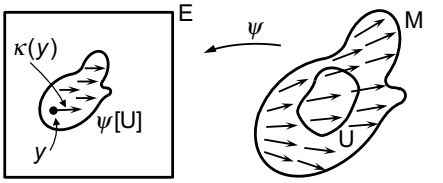
Now let x be that vector in E such that $(x; U, \psi)$ is in the equivalence class ξ. Then $x' = D(\psi' \cdot \varphi \cdot \psi^{-1})(\psi)(p))(x)$ is some vector in E'. We claim, first, that this x' is independent of the choice of the chart (U, ψ). [Indeed, given another, $\hat{U}, \hat{\psi}$, we must replace x by $\hat{x}$ such that $(\hat{x}; \hat{U}, \hat{\psi}) \approx (x; U, \psi)$. That is, we have $\hat{x} = D(\hat{\psi} \cdot \psi^{-1})(\psi(p))(x)$, which, by the chain rule, implies immediately that x' is unchanged.] Second, we claim that, if the chart U', ψ' on M' is changed to $\hat{U}', \hat{\psi'^{-1}}$, then x' is changed to $\hat{x'} = D(\hat{\psi'} \cdot \psi'^{-1})(\psi'(p))(x')$ (again, by the chain rule and the defining equation for x'). That is to say, we have that $(\hat{x'}; \hat{U}', \hat{\psi'}) \approx (x'; U', \psi')$. Thus, we obtain an equivalence class of pairs for p' in M'. That is, we obtain a tangent vector ξ' at p'. [Note that there is only one idea, and one kind of argument, in this business, which one keeps using over and over.]

Example. Regard the reals as a manifold. Then, using the obvious chart on this manifold, we may identify a tangent vector theorem with a real number. Consider now a C^p curve γ in manifold M, with $\gamma(0) = p$. By the construction above, this C^p mapping γ of manifolds takes the tangent vector "1" at 0 in $\mathbb{R}$ to some tangent vector at p in M. This tangent vector is, of course, what we on page 35 called the tangent vector to the curve γ.

Examples. Keep in force the first two sentences of the previous example. Let f be a C^p scalar field on manifold M, and fix a point p of M. Given any tangent vector ξ at p in M, we obtain, by the construction above, since f is a C^p mapping of manifolds, a tangent vector at $f(p)$ in R. We may identify this latter whit a real number. Thus, from f and ξ we obtain a real number. This number is of course what we called the directional derivative on page 35.

It is clear from these two examples why we kept using the same argument: We were doing exactly the same thing.

The final topic to be discussed in this section is that of tangent vector fields.



Let M be a C^p manifold ($p \geq 1$). A *tangent vector field* (or vector field) on M is a mapping which assigns, to each point p of M, a tangent vector at p in M. Thus, if ξ is a tangent vector field on M, then, for each point p in M, $\xi(p)$ is an element of the tangent space at p. If one represents a tangent vector at p as an arrow drawn on M at p, then a tangent vector field would be a field of arrows all over M.

As with scalar fields, the fields themselves are less interesting than the smooth ones. Let ξ be any tangent vector field on C^p M, and let U, ψ be any chart. Then, for any point y of the open subset $\psi[U]$ of E, $\psi^{-1}(y)$ is some point of M,. whence $\xi(\psi^{-1}(y))$ is a tangent vector at $\psi^{-1}(y)$. Let x be that (unique) vector in E such that $(x; U, \psi)$ is in the equivalence class $\gamma(\psi^{-1}(y))$. Thus, we have constructed a mapping κ from the open subset $\psi[U]$ of E to E (y goes to $x = \kappa(y)$). Our tangent vector field ξ on M is said to be C^{p-1} if this mapping κ is C^{p-1} for every admissible chart U, ψ on M. [That is, we "pull the vector field back to E by a chart, where we know what smoothness means".]

Why do we only define C^{p-1} vector fields on C^p manifold, and not C^p fields, or even C^q fields for arbitrary q? One could certainly make such a definition, but the difficulty would be that, unless $q \leq p - 1$, the only thing satisfying that definition would be the zero vector field. The reason for this difficulty is that, for a C^p manifold, the chart-maps are only C^p related. but the equivalence relation which defines tangent vectors has in it a derivative. Hence, given a vector field on C^p M, and a chart U, ψ on M such that the "pull-back" of the field to $\psi[U]$ looks, say C^p, then, for a second chart C^p – but not C^{p+1} – related to this one, the corresponding pull-back will not look C^p. But on a C^p manifold one must admit all C^p−compatible charts (even those which are only C^p – but not C^{p+1}−related to others). Hence, given a candidate for a C^p vector field on a C^p manifold (i.e., the field looks C^p in some charts), one expects to be able to find other admissible charts in which that candidate does not look C^p. The same question does not arise for scalar fields, for there there is no "equivalence relation involving one derivative". This behaviour is perhaps not unexpected, since a tangent vector already "looks at things to first order in M", i.e., a tangent vector has already within it "one derivative". Of course, if ξ is a C^{p-1} vector field on C^p M, and if $q \leq p$, then we may also regard M as a C^q manifold, whence this same ξ will be a C^{q-1} field theorem.

We now have a set of things to study: The set of tangent vector fields on M. As usual, the study consists of finding what structure there is on this set. Let ξ and η be C^{p-1} tangent vector fields on the C^p manifold M. Then, for each point p of M, $\xi(p)$ and $\eta(p)$ are tangent vectors at p, whence their sum, $\xi(p) + \eta(p)$, is well-defined, and is also a tangent vector at p. repeating for each p, we obtain a new tangent vector field on M, which we write $\xi + \eta$. We claim that this $\xi + \eta$ is in fact C^{p-1} (immediate, since the sum of the pull-backs of ξ and η via a chart is the pull-back of the sum, on since, by the third example on page 18, the sum of C^p mappings on Banach spaces is C^p). Further, for ξ a C^{p-1} vector field on M, and a real number, the vector field with action $(a\xi)(p) = a\xi(p)$ is also C^{p-1}. Thus, the set of C^{p-1} tangent vector fields on M has the structure of a vector space. Next, let ξ be a C^{p-1} vector field, and f a C^{p-1} scalar field, on M. Then action $(f\xi)(p) = f(p)\xi(p)$ defines another vector field on M, $f\xi$. This $f\xi$ is also C^{p-1} (since multiplication of C^{p-1} mappings of Banach spaces by C^{p-1} functions yields C^{p-1} mappings). That is, we can multiply vector fields by scalar fields. It is immediate that facts true pointwise are true for fields, i.e., $(f + f')\xi = f\xi + f'\xi$, $f(\xi + \eta) = f\xi + f\eta$. In algebraic language, the vector fields are a module over the ring $\mathfrak{I}$.

13. Tensor Products

We have now discussed two examples of fields. We shall shortly begin our program of obtaining the most general type of tensor field which can exist on a manifold. As a prerequisite for this program, we need a certain construction for obtaining Banach spaces from Banach spaces, a construction we now introduce.

Let E and $\hat{E}$ be Banach spaces. Denote by G the collection of all expressions of the form $a_1\, x_1 \otimes \hat{x_1} + a_2\, x_2 \otimes \hat{x_2} + \ldots$ such that i) the expression consists of a countably infinite, or possibly finite, or possibly zero, number of terms, ii) each of $a_1, a_2, \ldots$ is a nonzero real number, each of $x_1, x_2, \ldots$ a nonzero element of E, and each of $\hat{x_1}, \hat{x_2} \ldots$ a nonzero element of $\hat{E}$, iii) for no distinct i and j does both $x_i = x_j$ and $\hat{X}_i = \hat{x_j}$ (though, of course, we could have $x_i = x_j$, provided $\hat{x_i} \neq \hat{x_j}$), iv) the sum of positive real numbers, $|a_1||x_1||\hat{x_1}| + |a_2||x_2||\hat{x_2}| + \ldots$ converges, and v) two such expressions which differ only in the order of terms are taken as representing the same element of G. [We are being a bit sloppy here. More precisely: Consider the collection of all pairs, (S, λ), where S is a set whose cardinally is at most countable, and λ is a mapping from S to the product set $\mathbb{R} \times E \times \hat{E}$, subject to appropriate conditions. Write $(S, \lambda) \approx (S', \lambda')$ is there exists a one-to-one, onto mapping ρ from S to S' such that $\lambda' \cdot \rho = \lambda$. This is an equivalence relation: The set of equivalence classes is written G.]

We next wish to introduce structure on this set G which makes it into a Banach space. We first define addition. For $a_1\, x_1 \otimes \hat{x_1} + \ldots$ and $b_1\, y_1 \otimes \hat{y_1} + \ldots$ in G, define their sum as follows: First, "interlace terms", i.e., write $a_1\, x_1 \otimes \hat{x_1} + b_1\, y_1 \otimes \hat{y_1} + a_2\, x_2 \otimes \hat{x_2} + b_2\, y_2 \otimes \hat{y_2} + \ldots$ if then any two terms have the same "$x \otimes \hat{x}$", i.e., if both $a_i\, x \otimes \hat{x}$ and $b_j\, x \otimes \hat{x}$ appear, replace these two terms by $(a_i + b_i)\, x \otimes \hat{x}$ if $a_i + b_j$ is nonzero, and just remove the two terms if $a_i + b_j = 0$. Doing this for all repetitions, we obtain finally an element of G. [Examples: $3\, x \otimes \hat{x} + (-3)\, x \otimes \hat{x} = 0$; $2\, x \otimes \hat{x} + 1\, (-2x) \otimes \hat{x}$ is just itself, for there are no repetitions.] Next, we define multiplication by real numbers as follows; $a(a_1\, x_1 \otimes \hat{x} + a_2\, x_2 \otimes \hat{x_2} + \ldots) = (aa_1)\, x_1 \otimes \hat{x_1} + (aa_2)\, x_2 \otimes \hat{x_2} + \ldots$ (for $a \neq 0$; if $a = 0$, replace the right side by 0). This set G, with these two operations, is a vector space. [Think of the set of all expressions of the form "$x \otimes \hat{x}$", with $x \neq 0$ and $\hat{x} \neq 0$, as a "basis for G, so every vector in G can be written as a linear combination of these basis vectors". The quotes in the sentence above would be unnecessary if we had considered only finite expressions.]

Thus, we so far have constructed a vector space G. We next define a norm on this vector space as follows; $|a_1\, x_1 \otimes \hat{x_1} + \ldots| = |a_1||x_1||\hat{x_1}| + \ldots$ (noting condition iv) above). This is indeed a norm on G. [Clearly, $|a_1 x + 1 \otimes \hat{x_1} + \ldots| \geq 0$, and for equality, by ii), we must have no terms in the expression on the left. The third condition for a norm is obvious. For the second, note that in the operation of addition, one only combines or eliminates terms, and these combinations or eliminations serve only to decrease the norm of the sum of the norms.]

Thus, we now have a vector space with norm. We claim finally that this is complete, i.e., that we have a Banach. [Sketch of proof: Consider a Cauchy sequence in G, $a_1\, x_1 \otimes \hat{x_1} + \ldots, a'_1 x'_1 \otimes \hat{x'}_1 + \ldots$. Consider any expression $x \otimes \hat{x}$ which appears in at least one of the expressions representing the elements of this Cauchy sequence. Consider the sequence $r - 1, r - 2, \ldots$ of reals, where r_i is the coefficient of $x \otimes \hat{x}$ in the i^{th} expression (or zero, if $x \otimes \hat{x}$ does not appear in the i^{th} expression). Since our sequence in G is Cauchy, this sequence of real numbers must be Cauchy, whence it converges to some number r. Repeating for each choice (only a countable number need be tried) of $x \otimes \hat{x}$, we obtain a countable collection of $x \otimes \hat{x}$'s, each with a number r. Forgetting the $x \otimes \hat{x}$'s whose r's are zero, put the rest together in the expression $r\, x \otimes \hat{x} + r'\, x' \otimes \hat{x'} + \ldots$. But this expression satisfies all the conditions for membership in G. This is our candidate,

then, for the element of G to which our original Cauchy sequence converges. There only remains, therefore, the (completely standard) check of convergence.]

So, we obtain a Banach space G. This G, unfortunately, is not quite the thing we are looking for. What we would really like to have be true in G is that operation $\otimes$ be distributive, and that "numerical factors can be brought inside and applied to the x's without changing the element". That is to say, we would like to have the following four equations in G be true:

$$\begin{aligned}
&x \otimes (\hat{x'} + \hat{x'}) - x \otimes \hat{x} - x \otimes \hat{x'} = 0\\
&(x + x') \otimes \hat{x} - x \otimes \hat{x} - x' \otimes \hat{x} = 0\\
&a\,x \otimes \hat{x} - (ax) \otimes \hat{x} = 0\\
&a\,x \otimes \hat{x} - x \otimes (a\hat{x}) = 0
\end{aligned}$$

for all choices of the symbols which appear (where, of course, "$x \otimes \hat{x}$" means "$1\,x \otimes \hat{x}$", and "$-x \otimes \hat{x}$" means "$(-1)\,x \otimes \hat{x}$). But these four equations are not true in general in G: We want, therefore, to somehow "force" them to be true. This is accomplished as follows. First note that the intersection of any collection of subspaces of a Banach space is also a subspace. [Given vectors in the intersection, those vectors are in each subspace, whence any linear combination of them is in each subspace, whence that linear combination is in the intersection. So, the intersection is a vector subspace. Furthermore, the intersection, as the intersection of closed subset, is a closed subset. So, the intersection is a subspace.] Now, denote by H the intersection of all subspaces of G which contain all elements of the form of the left sides of the four equations above. This H is a subspace of G, and it consists precisely of "the things we would really rather have to zero in G". Finally, set $G/H = E \otimes \hat{E}$, a Banach space called the *tensor product* of Banach spaces E and $\hat{E}$.

We introduce the following notational conventions. Any expression satisfying the rules on page 39 defines an element of G, hence an element of G/H, hence an element of $E \otimes \hat{E}$. We shall allow ourselves to speak of such an expression as an element of $E \otimes \hat{E}$. Two elements of G whose difference is in the subspace H define the same element of G/H, and hence the same element of $E \otimes \hat{E}$. We shall represent this relationship by simply writing an equality sign between the elements (meaning "equal when regarded as elements of $E \otimes \hat{E}$"). Finally, we allow ourselves to write $0\,x \otimes \hat{x}$, a $0\,x \otimes \hat{x}$, a $0 \otimes \hat{x}$, and a $x \otimes \hat{0}$, these all being just other expression for 0 in $E \otimes \hat{E}$. With these conventions, then, the four formulae on the previous page are just true (in $E \otimes \hat{E}$). The only thing one has to be careful about, with these conventions, is the norm in $E \otimes \hat{E}$. It is not true, e.g., that the norm of $a\,x \otimes \hat{x} + a'\,x \otimes \hat{x'}$ in $E \otimes \hat{E}$ is necessarily $|a|\,|x|\,|\hat{x}| + |a'|\,|x'|\,|\hat{x'}|$ (for, e.g., the norm in $E \otimes \hat{E}$ of $x \otimes \hat{x} + (-x) \otimes \hat{x}$ is zero, since this is the zero element). Rather, the norm of such an expression is the greatest lower bound of all $|a_1|\,|x_1|\,|x\hat{x}_1| + \ldots$ with $a\,x \otimes \hat{x} + a'\,x' \otimes \hat{x'} = a_1\,x_1 \otimes \hat{x}_1 + a_2\,x_2 \otimes \hat{x}_2 + \ldots$ in $E \otimes \hat{E}$. Thus, we have in any case that, e.g., $|a\,x \otimes \hat{x} + a'\,x' \otimes \hat{x'}| \leq |a|\,|x|\,|\hat{x}| + |a'|\,|x'|\,|\hat{x'}|$. [These two remarks are just restatements of the definition of the norm in a quotient space of Banach spaces.]

The tensor product is one of those things whose definition, for some reason, is rather more complicated than the thing itself. We remark that one can, in a similar way, define the tensor product of three, or of any finite number, of Banach spaces.

We adopt the following additional convention. Let E be a Banach space, and $x_1, x_2, \ldots$ elements of E. The expression $x_1 + x_2 + \ldots$ is said to *converge* (absolutely) if the sum of real numbers, $|x_1| + |x_2| + \ldots$ converges. Then the sequence in E with n^{th} term $y_n = x_1 + \ldots + x_n$ is Cauchy (for, $n' > n, |y_{n'} - y_n| \leq |x_{n+1}| + \ldots + |x_{n'}|$), whence it converges to some element y of E. This y is called the *sum* of the x's, written $y = x_1 + x_2 + \ldots$. Note that these conventions are consistent with those we already have for infinite sums in tensor products.

We now give some examples of tensor products and their properties.

Example. Let E and $\hat{E}$ be finite-dimensional Banach spaces, of dimensions n and $\hat{n}$, respectively. We "find" their tensor product. Let $x_1, \ldots, x_n$ be a basis for E, and $\hat{x}_1, \ldots, \hat{x}_{\hat{n}}$

a basis for $\hat{E}$. Then, given any $n \times \hat{n}$ matrix $a_{i\hat{i}} (i = 1, \ldots, n; \hat{i} = 1, \ldots, \hat{n})$, we can obtain an element, $\Sigma_{i\hat{i}} a_{i\hat{i}} x_i \otimes \hat{x}_{\hat{i}}$, of $E \otimes \hat{E}$. Conversely, given element $x \otimes \hat{x}$ of $E \otimes \hat{E}$, we have, expanding each of x and $\hat{x}$ in terms of the bases – $x = \Sigma r_i x_i$ and $\hat{x} = \Sigma \hat{r}_i \hat{x}_i$ – that $x \otimes \hat{x}$ is of the form above, with $a_{i\hat{i}} = r_i \hat{r}_{\hat{i}}$. Hence, every element of $E \otimes \hat{E}$ consisting of a finite sum is of this form. Hence, so is every element of $E \otimes \hat{E}$ of finite sums, and since limits of elements of the form, $\Sigma a_{i\hat{i}}\, x_i \otimes \hat{x}_{\hat{i}}$, are also of this form). Finally, note that two elements of $E \otimes \hat{E}$ written in our canonical form, $\Sigma a_{i\hat{i}}\, x_i \otimes \hat{x}_{\hat{i}}$ and $\Sigma a'_{i\hat{i}} x_i \otimes \hat{x}_{\hat{i}'}$, are equal in $E \otimes \hat{E}$ when and only when $a_{i\hat{i}} = a'_{i\hat{i}}$ for all i and $\hat{i}$. We conclude, therefore, that $E \otimes \hat{E}$ is, as a vector space, the same as the vector space of $n \times \hat{n}$ matrices. In particular, $E \otimes \hat{E}$ is $n\hat{n}-$dimensional. [Note that $E \times \hat{E}$ is $(n + \hat{n})-$dimensional.]

Example. Let E be a Banach space. We "find" the tensor product $E \otimes \mathbb{R}$. Given any element, $a_1\, x_1 \otimes b_1 + a_2\, x_2 \otimes b_2 + \ldots$ of $E \otimes \mathbb{R}$ (where the a's and b's are reals, and the x's are in E), this is equal to $(a_1, b_1)x_1 \otimes 1 + (a_2 b_2)x_2 \otimes 1 + \ldots$ and hence to $(a_1 b_1 x_1) \otimes 1 + \ldots$). But this, in turn, is equal to $(a_1 b_1 x_2 + a_2 b_2 x_2 + \ldots) \otimes 1$ (for, each n, $(a_1 b_1 x_1 + \ldots + a_n b_n x_n) \otimes 1 - (a_1 b_1 x_1) \otimes 1 - \ldots - (a_n b_n x_n) \otimes 1$ is in H, whence, since H is closed, so is the limit as n goes to infinity). Thus, every element of $E \otimes R$ can be written in the form $x \otimes 1$, for x in E and, conversely, every x in E defines an element $x \otimes 1$ of $E \otimes \mathbb{R}$.This one-to-one, onto linear mapping from $E \otimes \mathbb{R}$ to E is norm-decreasing, since $|x \otimes 1| \leq |x|$, and so is an isomorphism of Banach spaces, by the open mapping theorem. [In fact, $|x \otimes 1| = |x|$.] Thus, $E \otimes \mathbb{R}$ is isomorphic with E.

Example. Let E, $\hat{E}$ and F be Banach spaces. We find an isomorphism from $\mathcal{L}(E \otimes \hat{E}; F)$ to $\mathcal{L}(E, \hat{E}; F)$. First, let α be in $\mathcal{L}(E \otimes \hat{E}; F)$. We associate, with this α element β of $\mathcal{L}(E\hat{E}; F)$, with action $\beta(x, \hat{x}) = \alpha(x \otimes \hat{x}$ (noting that this β is indeed bilinear; and that it is bounded, since α is bounded, and since $|x \otimes \hat{x}| \leq |x|\,|\hat{x}|$). For the converse, let β be in $\mathcal{L}(E, \hat{E}; F)$. Then, for $a_1\, x_1 \otimes \hat{x}_1 + \ldots$ in $E \otimes \hat{E}$, set $\alpha(a_1\, x_1 \otimes \hat{x}_1 + \ldots) = a_1 \beta(x_1, \hat{x}_1) + a_2\, \beta(x_2, \hat{x}_2) + \ldots$. [The sum on the right converges. We have $\Sigma |a_i\, \beta(x_i, \hat{x}_i)| \leq \Sigma |\beta| |a_i|\, |x_i|\, |\hat{x}_i| = |\beta| \Sigma |a_i|\, |x_i|\, |\hat{x}_i|$, while this last sum is finite by membership in the tensor products.] Thus, we have a one-to-one, onto linear mapping from $\mathcal{L}(E \otimes \hat{E}; F)$ to $\mathcal{L}(E, \hat{E}; F)$. Furthermore, this mapping is norm-decreasing, for, for α and β related as above and z in $E \otimes \hat{E}$, we have, by our last calculation above, that $|\alpha(z)| \leq |\beta|\, |z|$, whence $|\alpha| \leq |\beta|$. Hence, it is an isomorphism of Banach spaces (open mapping).

Example. Let E, F, and G be Banach spaces. We first obtain a mapping from $\mathcal{L}(E; F) \otimes G$ to $\mathcal{L}(E; F \otimes G)$. Let $\alpha_1 \otimes z_1 + \alpha_2 \otimes z_2 + \ldots$ be in the former (so the α's are in $\mathcal{L}(E; F)$, the z's in G). Associate with this the element of $(E; F \otimes G)$ which sends x in E to $\alpha_1(x) \otimes z_1 + \ldots$ (noting that this last satisfies the condition for membership in $F \otimes G$, since $\Sigma |\alpha_i(x)|\, |z_i| \leq |x| \Sigma |\alpha_i|\, |z_i|$ is finite). Call this mapping (clearly linear) from $\mathcal{L}(E; F) \otimes G$ to $\mathcal{L}(E; F \otimes G)$ ψ. For $\alpha_i \otimes z_1 + \ldots$ in $\mathcal{L}(E; F) \otimes G$, and x in E, $\psi(\alpha_1 \otimes z_1 + \ldots)(x) = |\alpha_1(x) \otimes z_1 + \ldots| \leq |\alpha_1(x)|\, |z_1| + \ldots \leq |x|(|\alpha_1|\, |z_1| + \ldots)$. Hence, ψ is a bounded linear mapping, with norm less than or equal to one. It seems likely that this ψ is also one-to-one, and that its image is a closed subspace of $\mathcal{L}(E; F \otimes G)$.

We show that, however, ψ is not onto, in general. Let $F = \mathbb{R}$ and $G = E$, so ψ is a bounded linear mapping from $\mathcal{L}(E; R) \otimes E$ to $\mathcal{L}(E; E)$. If the identity in $\mathcal{L}(E; E)$ were in the domain of ψ, say were $\psi(\alpha_1 \otimes x_1 + \ldots)$, then we would have that every element of E is a (possibly infinite) linear combination of the x_i's. Thus, we have only to find a Banach space E having no countable subset linear combinations of which give every element of E. To this end, let S be any set, and let E be the Banach space of real-valued, bounded functions on S, with norm the least upper bound of the function. "The larger S, the larger E." Clearly, by choosing S to have sufficiently large cardinality, the corresponding E will have the property above. [In fact, ψ will not be onto for any infinite-dimensional E.]

For finite-dimensional E, F and G, the map of the first paragraph of this example will be an isomorphic of Banach spaces.

We remark, finally, that there is a universal definition of the tensor product, as follows. Let E and F be Banach spaces. A tensor product of E and F consists of a Banach space G,

together with a bounded bilinear mapping ψ from E, F to G such that, given any Banach space G' and any bounded bilinear mapping ψ' from E, F to G', there is a unique bounded linear mapping κ from G to G' with $\psi' = \kappa \cdot \psi$. One shows, first, that, if a tensor product so defined exists, then it is unique. Then, one shows that what we have constructed is it.

14. Tensor Spaces

Our goal is to find, and learn to manipulate, the various kinds of fields which can exist on a manifold. Let, then, M be a manifold based on Banach space E. We can divide our program into two parts: i) Obtain the various Banach spaces that can be constructed from E, and unravel their structure, ii) Obtain, from each such space, a corresponding type of field on M, and carry over what structure one can from the Banach spaces to the fields. The first is essentially an algebraic problem, the second a differential one. One might suspect, therefore, that the first will be the easier (applying here the adage "algebraic is easier than differential"). This seems, however, not to be the case, and, indeed, the first problem has not, as far as I am aware, been put into a state that I would call totally satisfactory. In this section, we state the first problem; in the next, we describe some possible lines toward its solution.

Fix, once and for all, a Banach space E. Consider i) the Banach spaces E and $\mathbb{R}$, and ii) the constructions of multilinear mappings "$\mathcal{L}(,\ldots,;)$" and tensor products "$\otimes$". We consider now all Banach spaces obtained by applying the constructions ii) to the Banach spaces i), or the constructions to the Banach spaces so obtained, etc. For example, one such would be $\mathcal{L}(E\otimes(\mathbb{R}\otimes\mathcal{L}(E;R))$, $\mathcal{L}(E;E\otimes E)$; $\mathbb{R}\otimes\mathcal{L}(\mathcal{L}(\mathcal{L}(E;\mathbb{R});\mathbb{R});E))\otimes E$. We call these Banach spaces the *tensor spaces* (over E), and the elements of these Banach spaces *tensors* (over E). [If we wish to give more detail, an element of tensor space A will be called and $A-$tensor.] These definitions are inadequate for two reasons. First, the "applying constructions" business is not precise. Second, it is not made clear that, when we speak of a tensor space, we intend to refer, not only to the Banach space itself, but also to the sequence of constructions by which it arises from E. Both of these inadequacies could be resolved by proceeding as follows. We could define a tensor space as a finite sequence of symbols chosen from "E", "$\mathbb{R}$", "$\mathcal{L}$", "$\otimes$", "," ";", "(", and "), subjects to certain rules. One would then note that, if E is a Banach space, then each finite sequence gives rise to a certain Banach space. With these last five sentences as an appendix to the definition, there should be no ambiguity.

There is of course an enormous amount of structure on the tensor spaces and tensors over E. For example:

1. Natural operations on tensors. The sum of two $A-$tensors is well-defined (since tensor space A is a Banach space) as is the limit of a sequence of $A-$tensors. The tensor product of an $A-$tensor α and a $B-$tensor β can be taken, and $\alpha\otimes\beta$ is an $(A\otimes B)-$tensor. For α and $A-$tensor and τ a $\mathcal{L}(A;B)-$tensor, $\tau(\alpha)$ is a $B-$tensor. For τ a $\mathcal{L}(A;B)-$tensor and σ a $\mathcal{L}(B;C)-$tensor, $\sigma\cdot\tau$ is a $\mathcal{L}(A;C)-$tensor. With any $(A\otimes\mathbb{R})-$tensor we may associate an $A-$tensor (first example, page 41).

2. Preferred tensors. Some tensor spaces have proffered elements. For example, the identity in $\mathcal{L}(A;A)$ is a "preferred" $\mathcal{L}(A;A)-$tensor. So is the element of $\mathcal{L}(A,B;A\otimes B)$ which takes the tensor product; the element of $\mathcal{L}((A;B)\otimes C;\mathcal{L}(A;B\otimes C))$ of the last example on page 41; the element of $\mathcal{L}(\mathcal{L}(\mathbb{R};A);A)$ which sends the element α of $\mathcal{L}\mathbb{R};A)$ to the element $\alpha(1)$ of A; the element "1" of $\mathbb{R}$.

3. Naturally isomorphic tensor spaces. Some pairs of tensor spaces are just different ways of writing essentially the same thing. For example, for A a tensor space, $\mathcal{L}(\mathbb{R};A)$ and A are naturally isomorphic; $\mathbb{R}\otimes A$ and A are naturally isomorphic. For A, B, and C tensor spaces, $\mathcal{L}(A,B;C)$ and $\mathcal{L}(A\otimes B;C)$ are naturally isomorphic.

4. Tensor spaces natural subspaces of others. For A any tensor space, A is a natural subspace of $\mathcal{L}(\mathcal{L}(A;\mathbb{R});\mathbb{R})$. Thus, any $A-$tensor can also be regarded as a $\mathcal{L}(\mathcal{L}(A;\mathcal{R});)-$tensor,but not conversely. Similarly, $E\otimes\mathcal{L}(E;R)$ is a natural subspace of $\mathcal{L}(E;E)$.

5. Tensor spaces natural quotients of others. For example, for A and B tensor spaces, E may be regarded as a quotient of the tensor space $A\otimes\mathcal{L}(A;B)$ (namely, the quotient by

the kernel of the natural mapping from $A \otimes \mathcal{L}(A; B)$ to B).

Problem: Organize this situation. What tensor spaces are available? Which are isomorphic to which? Which natural subspaces or natural quotients? What operations are available on tensors? Which tensor spaces have preferred elements, and what are they? One would like to cast this subject into some manageable form, in which all elementary facts look elementary, in which complicated calculations can be performed with relative ease, in which one doesn't have to continually go back and prove new things about Banach-space operations. There are of course two halves to the problem: i) make precise terms such as "natural" and "preferred", and ii) organize what is available.

15. Natural Tensors

In the previous section we posed a rather vaguely-stated problem (in which, indeed, part of the problem is to eliminate the vagueness). We now sketch a few notions which, we suggest, offer a possible line to a solution. Our basic claim is that one can make a definition which, on the one hand, gives reasonable meaning to such words as "natural" and "preferred", and, on the other, encompasses everything in which we are interested – operations, preferred elements, isomorphic tensor spaces, etc.

We first note the following facts. Let A, A', B, and B' be Banach spaces (note necessarily tensor spaces). Let α be an isomorphism from A to A', and β an isomorphism from B to B'. We define a corresponding isomorphism τ from $\mathcal{L}(A;B)$ to $\mathcal{L}(A';B')$ as follows: For κ in $\mathcal{L}(A;B)$, $\tau(\kappa)$ is the element of $\mathcal{L}(A';B')$ given by $\beta\cdot\kappa\cdot\alpha^{-1}$. That is, for x' in A', $\tau(\kappa)(x') = \beta\cdot\tau(\alpha^{-1}(x'))$. Similarly, given isomorphisms from A_1 to A'_1, etc, to A'_n, and from B to B', we obtain an isomorphism from $\mathcal{L}(A_1,\ldots,A_n;B)$ to $\mathcal{L}(A'_,\ldots,A'_n;B')$. [Note that we have no freedom in writing this definition.] Next, again let A and B be Banach spaces, and again let α and β be isomorphisms from A and B to Banach spaces A' and B', respectively. Then we can define an isomorphism from $A\otimes B$ to $A'\otimes B'$ as follows: For $x_1\otimes y_1+x_2\otimes+\ldots$ in $A\otimes B$, $\tau(x_1\otimes y_1+\ldots) = \alpha(x_1)\otimes\beta(y_1)+\alpha(x_2)\otimes\beta(y_2)+\ldots$ (noting that the sum on the right converges, by boundedness of α and β, and by the fact that the sum on the left converges). Of course, these are all isomorphisms: Their inverses are obtained by the same constructions from the inverses of α and β. Thus, we can extend isomorphisms given on some Banach spaces to isomorphisms on Banach spaces constructed (by our two constructions) from these.

Now fix Banach space E. Let ι be any isomorphism from E to E, and also let ι be the identity isomorphism from $\mathbb{R}$ to $\mathbb{R}$. [This use of one letter for two things greatly conserves symbols, and causes no confusion.] By the paragraph above we obtain, given any Banach space constructed from E's and $\mathbb{R}$'s by taking multilinear mapping and tensor products, an isomorphism from this Banach space to itself. Hence, on any Banach space constructed from these by our two operations, we obtain also an isomorphism. Continuing in this way, we extend the action of ι from just E and $\mathbb{R}$ to all the tensor spaces over E.

In more detail, then, the situation is as follows. Let A and B be tensor spaces over E, and suppose that we have obtained the action of ι on A and B. Then the action of ι on $\mathcal{L}(A;B)$ is as follows: For α in $\mathcal{L}(A;B)$, $\iota(\alpha) = \iota\cdot\alpha\cdot\iota^{-1}$. Similarly, the action of ι on $A\otimes B$ is: For $x_1\otimes y_1+\ldots$ in $A\otimes B$, $\iota(x_1\otimes y_1+\ldots) = \iota(x_1)\otimes\iota(y_1)+\ldots$

To summarize, any given isomorphism on E extends in a "natural" way to isomorphisms on all the tensor space constructed from E.

The key definition is this: element α of tensor space A is called a *natural tensor* (or natural $A-$tensor) if $\iota(\alpha) = \alpha$ for every ι. In other words, the natural tensors remain invariant under any isomorphism induced from one on E, i.e., they in some sense "just exist, no matter what E is doing or like". [Those familiar with category theory will recognize this definition as a thinly disguised version of a natural transformation.]

Example. Each real number is a natural tensor (since ι is the identity on $\mathbb{R}$).

Example. For any tensor space A, the identity in $\mathcal{L}(A;A)$ is a natural tensor. Indeed, for τ in $\mathcal{L}(A;A)$, and for x in A, we have $\iota(\tau)$ defined by $\iota(\iota(\tau)(\iota^{-1}(x))) = \tau(x)$. But, for τ the identity, this is clearly satisfied by $\iota(\tau) = \tau$.

Example. Let A and B be any tensor spaces. then, clearly, the zero element of $\mathcal{L}(A;B)$ is a natural tensor.

Example Let A be any tensor space, and let τ be the element of $\mathcal{L}(\mathcal{L}(\mathbb{R};A);A)$ which sends the element of α of $\mathcal{L}(\mathbb{R};A)$ to $\alpha(1)$. Now, for any α in $\mathcal{L}(\mathbb{R};A)$, $\iota(\alpha)$ is that element of $\mathcal{L}(\mathbb{R};A)$ with action $\iota(\alpha)(a) = \iota(\alpha(a))$; for any τ in $\mathcal{L}(\mathcal{L}(\mathbb{R};A);A)$, $\iota(\tau)$ has action $\iota(\tau)(\alpha) = \iota(\tau(\iota^{-1}(\alpha))$. Now let this τ be that above. Then the right side of the last equation

is $\iota((\iota^{-1}(\alpha))(1))$, which, by the action of ι on $\mathcal{L}(\mathbb{R};A)$, is $\iota(\iota^{-1}(\alpha(1))$, which equals $\alpha(1)$. Thus, $\iota(\tau)(\alpha) = \alpha(1)$. But this right side is just $\tau(\alpha)$. So, $\iota(\tau)(\alpha) = \tau(\alpha)$ for every α, i.e., $\iota(\tau) = \tau$. We have shown, therefore, that this τ is a natural tensor. Similarly, the inverse of this τ, an element of $\mathcal{L}(A;\mathcal{L}(\mathbb{R};A))$, is a natural tensor.

Example. Let A, B and C be tensor spaces. Denote by τ the element of $\mathcal{L}(\mathcal{L}(A,B;C);\mathcal{L}(A;\mathcal{L}(B;C)))$ which sends μ in $\mathcal{L}(A,B;C)$ to the element of $\mathcal{L}(A;\mathcal{L}(B;C))$ which sends x in A to the element of $\mathcal{L}(B;C)$ which sends y in B to $\mu(x,y)$ in C. We show that this τ is a natural tensor. First note that, for μ in $\mathcal{L}(A,B;C)$, $\iota(\mu)$ has action $\iota(\mu)(x,y) = \iota(\mu(\iota^{-1}(x),\iota^{-1}(y)))$ (with x in A and y in B). Also, for ν in $\mathcal{L}(A;\mathcal{L}(B;C))$, $\iota(\nu)$, has action $((\iota(\nu))(x))(y) = \iota((\nu(\iota^{-1}(x)))(\iota^{-1}(y)))$. Finally, for any τ in $\mathcal{L}(\mathcal{L}(A,B;C);\mathcal{L}(A;\mathcal{L}(B;C)))$, $\iota(\tau)$ has action $\iota(\tau)(\mu) = \iota(\tau(\iota^{-1}(\mu)))$. Applying ι^{-1} to this last equation, we have $\iota^{-1}(\iota(\tau)(\mu)) = \tau(\iota^{-1}(\mu))$. Now let τ be that tensor given in the second sentence of this example. Each side of this last equation is an element of $\mathcal{L}(A;\mathcal{L}(B;C))$: Applying each side to x in A (to obtain an element of $\mathcal{L}(B;C)$) and then to y in B (to obtain an element of C), we have $\iota^{-1}(\iota(\tau)(\mu))(x)(y) = \tau(\iota^{-1}(\mu))(x)(y)$. Using the defining equation for our particular τ on the right, this right side is $(\iota^{-1}(\mu))(x,y)$. Using now the action of ι on $\mathcal{L}(A,B;C)$, this in turn is $\iota^{-1}(\mu(\iota(x),\iota(y)))$. Using again the definition of τ, this in turn is $\iota^{-1}(\tau(\mu)(\iota(x))(\iota(y)))$. Thus, we have so far $\iota^{-1}(\iota(\tau)(\mu))(x)(y) = \iota^{-1}(\tau(\mu)(\iota(x))(\iota(y)))$. Using now on the left the action of ι on $\mathcal{L}(A;\mathcal{L}(B;C))$, this left side is $\iota^{-1}(\iota(\tau)(\mu)(\iota(x))(\iota)y)))$. Thus, we have $\iota^{-1}(\iota(\tau)(\mu)(\iota(x))(\iota(y))) = \iota^{-1}(\tau(\mu)(\iota(x))(\iota(y)))$. Since ι is an isomorphism on C, this implies $\iota(\tau)(\mu)(\iota(x))(\iota(y)) = \tau(\mu)(\iota(x))(\iota(y))$. Since x and y are arbitrary, this implies $\iota(\tau)(\mu) = \tau(\mu)$. Since μ is arbitrary, this implies $\iota(\tau) = \tau$. Thus, our τ is indeed a natural tensor. Similarly, the inverse of τ is a natural tensor. Similarly, with "A,B" replaced by more Banach spaces in $\mathcal{L}(A,B;C)$.

Example. Let A,B and C be tensor spaces. Denote by τ the element of $\mathcal{L}(\mathcal{L}(A\otimes B;C),\mathcal{L}(A,B;C))$ which sends μ in $\mathcal{L}(A\otimes B;C)$ to that element $\nu = \tau(\mu)$ of $\mathcal{L}(A,B;C)$ which sends x in A and y in B to $\nu(x,y) = \mu(x\otimes y)$. We show that this τ is a natural tensor. For μ in $\mathcal{L}(A\otimes B;C)$, $\iota(\mu)$ has action $\iota(\mu)(z) = \iota(\mu(\iota^{-1}(z)))$ (for, of course, z in $A\otimes B$). For ν in $\mathcal{L}(A,B;C)$, $\iota(\nu)$ has action $\iota(\nu)(x,y) = \iota(\nu(\iota^{-1}(x),\iota^{-1}(y)))$. For any τ in $\mathcal{L}(\mathcal{L}(A\otimes B;C);\mathcal{L}(A,B;C))$, $\iota(\tau)$ has action $\iota(\tau)(\mu) = \iota(\tau(\iota^{-1}(\mu)))$. Applying ι^{-1} to this last equation, we have $\iota^{-1}(\iota(\tau)(\mu)) = \tau(\iota^{-1}(\mu))$. Now let τ be that particular tensor given above, and apply each side of this equation to (x,y), to obtain $\iota^{-1}(\iota(\tau)(\mu))(x,y) = \tau(\iota^{-1}(\mu))(x,y)$. By definition of τ, the right side is $(\iota^{-1}(\mu))(x\otimes y)$. By the action of ι on $\mathcal{L}(A\otimes B;C)$, this in turn is $iota^{-1}(\mu(\iota(x\otimes\iota(y)))$. By the action of τ on tensor products, this in turn is $\iota^{-1}(\mu(\iota(x)\otimes\iota(y))$. By definition of τ, this is $\iota^{-1}(\tau(\mu)(\iota(x),\iota(y)))$. Thus, we have so far that $\iota^{-1}(\iota(\tau)(\mu))(x,y) = \iota^{-1}(\tau(\mu))\iota(x),\iota(y)))$. Applying the action of ι on $\mathcal{L}(A,B;C)$ to the left side, this becomes $\iota^{-1}((\iota(\tau)(\mu))(\iota(x))(\iota(y))) = \iota^{-1}(\tau(\mu)(\iota(x),\iota(y)))$. Since ι is an isomorphism, this implies $(\iota(\tau)(\mu))(\iota(x))(\iota(y)) = \tau(\mu)(\iota(x),\iota(y))$. Since x and y are arbitrary, this implies $\iota(\tau)(\mu) = \tau(\mu)$. Since μ is arbitrary, this implies $\iota(\tau) = \tau$. Thus, τ is a natural tensor. Similarly for its inverse.

These little calculations are easier they might appear. All one must do is i) make sure that, in each step, one does something new, rather than retracing the previous step, and ii) be careful not to leave out any parentheses.

We now have the notion of a natural tensor, together with some examples. We next face two issues. First, we must make a case that essentially all structure of interest on the tensor spaces can be expressed in terms of these natural tensors. Second, we must find some way easier than that of the examples above to check naturality (for the situation would be hopeless if we had to go through all that agony for each natural tensor), and we must classify them all. We discuss these two issues in turn.

For the first issue, we begin with some definitions. For $A_1,\ldots,A_n$ and B tensor spaces, a natural $\mathcal{L}(A_1,\ldots,A_n;B)-$tensor τ will be called a *natural operation*. [We regard this τ as the operation which assigns to x_1 in $A_1,\ldots,x_n$ in A_n the tensor $\tau(x_1,\ldots,x_n)$ in tensor space B.] For A and B tensor spaces, a natural $\mathcal{L}(A,B)-$tensor for which there exists a natural $\mathcal{L}(B,A)-$tensor such that these two linear mappings are each others inverses will

be called a *natural isomorphism* (from A to B). When such a natural isomorphism exists, we say that A and B are *naturally isomorphic*. We could, similarly, define a proffered tensor in tensor space A as a natural $A-$tensor (but we will not, since we already have the term "natural tensor" for this purpose). Finally, we could define natural subspaces and natural quotients in the obvious way (but we will not, since we will not need these terms).

Of course, the mere introduction of these terms does not a case make. What we must now do is show, by means of a large collection of examples, that the various things one would wish intuitively to call a "natural operation" "natural isomorphism", "preferred element", etc. actually arises as such from the definitions above. Further, one would like to show that each thing named above actually has the intuitive connotations of that name. We emphasize that there is nothing to prove here: It is only a matter of eliciting conviction by means of examples. We postpone, for a moment, these examples.

For the second issue, we must discover some easy way to find all the natural tensors. To this end, we first make the following two observations: i) For τ any natural $\mathcal{L}(A_1, \ldots, A_n; B)-$tensor, and $\alpha_1, \ldots, \alpha_n$ any natural A_1-, ..., A_n-tensors, respectively, $\tau(\alpha_1, \ldots, \alpha_n)$ is a natural $B-$tensor. [Indeed, we have, by the action of ι on $\mathcal{L}(A_1, \ldots, A_n; B)$, that $\iota(\tau(\alpha_1, \ldots, \alpha_n)) = \iota(\tau)(\iota(\alpha_1), \ldots, \iota(\alpha_n))$. But, by naturality, the right side is just $\tau(\alpha_1, \ldots, \alpha_n)$.] ii) For μ any natural $\mathcal{L}(A, B)$ $A-$tensor and ν any natural $\mathcal{L}(B; C)$ $A-$tensor, $\nu \cdot \mu$ is a natural $\mathcal{L}(A; C)-$tensor (since $\iota(\nu \cdot \mu) = \iota(\nu) \cdot \iota(\mu) = \mu \cdot \mu$.). We next note that we already have, from the discussion above, the following six examples of natural tensors: 1. For any tensor space A, the zero tensor in A is natural. 2. Any real number is a natural $\mathbb{R}-$tensor. 3. For any tensor space A, the identity in $\mathcal{L}(A; A)$ is a natural $\mathcal{L}(A; A)-$tensor. 4. For any tensor space A, we have (fourth example on page 45) a natural tensor in $\mathcal{L}(\mathcal{L}(\mathbb{R}; A); A)$, and its inverse, a natural tensor in $\mathcal{L}(A; \mathcal{L}\mathbb{R}; A))$. 5. For any tensor spaces $A_1, \ldots, A_n$ and B, we have (fifth example on page 45) a natural tensor in $\mathcal{L}(\mathcal{L}(A_1, \ldots, A_n; B); \mathcal{L}(A_1, \ldots, A_{i-1}, A_{i+1}, \ldots A_n; \mathcal{L}(A_i; B))$, together with its inverse. 6. For any tensor spaces A, B, and C, we have (first example on page 46) a natural $\mathcal{L}(\mathcal{L}(A \otimes B; C); \mathcal{L}(A, B; C))-$tensor, and its inverse. Thus, we have already some examples of a few natural tensors, as well as some constructions which yield natural tensors from natural tensors. We now formulate

Conjecture. Every natural tensor is a sum of natural tensor obtained by applying the constructions i) –ii) above to the examples 1 – 6 above. If this conjecture were true, then one would have at least some control over the natural tensors, for one would have a relatively simple algorithm for obtaining them (rather then the more complicated algorithm "guess at one, and then go through all the ι's to check that guess if it is wrong, try another guess"). As its title suggests, I know of neither a proof nor a counterexample (although of course it may very well be resolved, somewhere in the literature). We can, however, at least support this conjecture by means of examples, i.e., by demonstrating that various natural tensors that come to mind can indeed be obtained by the rules laid down in the Conjecture. In my view, it would be of some interest to first resolve this conjecture, and then to begin using it (or some modification, if the conjecture should turn out to be false) to "find" all natural tensors.

Thus, we now have two issues, each of which turns to a certain extent (although, technically, in rather different ways) on examples. It is our intention, now, to support the discussion above by means of various examples. In each example, we shall find a natural tensor using the rules set forth in the conjecture (and thus support the conjecture), while at the same time many of the natural tensions we find will correspond intuitively to "natural operations", "natural isomorphisms", etc. (thus supporting the discussion of the first issue). In short, we intend each example below to serve two roles.

Example. Denote by τ the element of $\mathcal{L}(A, \mathcal{L}(A; B); B)$ with action $\tau(x, \alpha) = \alpha(x)$. We show that this τ is natural. By example 3, the identity in $\mathcal{L}(\mathcal{L}(A; B); \mathcal{L}(A; B))$ is natural. By example 5, we have a natural element of $\mathcal{L}(\mathcal{L}(\mathcal{L}(A; B); \mathcal{L}(A : B)); \mathcal{L}(A, \mathcal{L}(A; B); B))$. Applying the latter to the former, we obtain τ. [Thus, "application" is a natural operation on tensor spaces.]

Example. Denote by τ the element of $\mathcal{L}(A, \mathcal{L}(A : B) : B)$ with action $\tau(x, \alpha) = \alpha(x)$. We show that this τ is natural. By example 3, the identity in $\mathcal{L}(\mathcal{L}(A : B) : \mathcal{L}(A : B))$ is natural. By example 5 we have a natural of $\mathcal{L}(\mathcal{L}(A : B) : \mathcal{L}(A : B)) : (A, \mathcal{L}(A : B) : B))$. Applying the latter to the former, we obtain τ. [Thus, "application" is a natural operation on tensor spaces.]

Example. Denote by τ the element of $\mathcal{L}(\mathcal{L}(A; B), \mathcal{L}(B; C); \mathcal{L}(A; C))$ with action $\tau(\alpha, \beta) = \beta \cdot \alpha$. We show that this τ is natural. By the example above, we have a natural element of $\mathcal{L}(B; \mathcal{L}(B; C); C)$. Applying to this the natural element of $\mathcal{L}(\mathcal{L}(B, \mathcal{L}(B; C); C); \mathcal{L}(B; \mathcal{L}(\mathcal{L}(B; C); C)))$ (example 5), we obtain a natural element of $\mathcal{L}(B; \mathcal{L}(\mathcal{L}(B); C); C))$. By the example above, we also have a natural element of $\mathcal{L}(A, \mathcal{L}(A; B); B)$. Composing these two, we obtain a natural element of $\mathcal{L}(A, \mathcal{L}(A; B); \mathcal{L}(\mathcal{L}(B; C); C))$. Again applying the natural tensor of example 5 to this one, we obtain a natural element of $\mathcal{L}(A, \mathcal{L}(A; B), \mathcal{L}(\mathcal{L}(B; C); C)$. Finally, applying the natural tensor of example 5 once again, we obtain a natural element of $\mathcal{L}(\mathcal{L}(A; B), \mathcal{L}(B; C); \mathcal{L}(A; C))$. This is our τ. [Thus, "composition" is a natural operation on tensor spaces.]

Example. The tensor product of two natural tensors is natural. By example 5, we have a natural element of $\mathcal{L}(A \otimes B; A \otimes B)$. Composing with the natural tenor of element 6, we obtain a natural element τ of $\mathcal{L}(A, B; A \otimes B)$. Hence, given natural tensors in A and B, α and β, respectively, $\tau(\alpha, \beta)$ is a natural tensor (application of natural tensors to natural tensors).

Example. Denote by τ the element of $\mathcal{L}(R, A; A)$ with action $\tau(a, \alpha) = a\alpha$. We show that this τ is natural. By example 4, we have a natural element of $\mathcal{L}(A; \mathcal{L}(R; A))$. By example 5, we have a natural element of $\mathcal{L}(\mathcal{L}(A; \mathcal{L}(R; A)); \mathcal{L}(R, A; A))$. Applying the second to the first, we obtain τ, a natural element of $\mathcal{L}(R, A; A)$. [Thus, "scalar multiplication" is a natural operation on tensor spaces.]

Example. Denote by τ the element of $\mathcal{L}(A; R \otimes A)$ with action $\tau(\alpha) = 1 \otimes \alpha$. We show that this τ is natural. Applying the natural isomorphism (example 6) from $\mathcal{L}(R \otimes A; R \otimes A)$ to $\mathcal{L}(R, A; R \otimes A)$ to the identity in the former, we obtain a natural element of $\mathcal{L}(R, A; R \otimes A)$. Applying the natural isomorphism of example 5, we obtain a natural element of $\mathcal{L}(A; \mathcal{L}(R; R \otimes A))$. Composing this with the natural isomorphism (example 4) from $\mathcal{L}(R; R \otimes A)$ to $R \otimes A$, we obtain a natural element of $\mathcal{L}(A; R \otimes A)$. This is our τ.

Example Denote by τ the element of $\mathcal{L}(A; \mathcal{L}(\mathcal{L}(A; R); R))$ which sends α in A to the element of $\mathcal{L}(\mathcal{L}(A; R); R)$ which sends μ in $\mathcal{L}(A; R)$ to the number $\mu(\alpha)$. We show that this τ is natural. By the first example on page 46, we have a natural element of $\mathcal{L}(A, \mathcal{L}(A; R); R)$. Applying the natural isomorphism of example 5, we obtain our τ. [This is the operation of "inserting A into its double dual".]

Example. Denote by τ the element of $\mathcal{L}(A \otimes \mathcal{L}(A; B); B)$ with action $\tau(x_1 \otimes \mu_1 + \ldots) = \mu_1(x_1) + \ldots$. We show that this τ is natural. Applying the natural isomorphism from $\mathcal{L}(A, \mathcal{L}(A; B); B)$ to $\mathcal{L}(A \otimes \mathcal{L}(A; B); B)$ to the natural element of $\mathcal{L}(A, \mathcal{L}(A; B); B)$, we obtain our τ.

Example. Denote by τ the element of $\mathcal{L}(A \otimes B; B \otimes A)$ which "switches the order of factors". We show that this τ is natural. We have a natural isomorphism from $\mathcal{L}(A \otimes B; A \otimes B)$ to $\mathcal{L}(A, B; A \otimes B)$, and from this last to $\mathcal{L}(B, A; A \otimes B)$, and from this last to $\mathcal{L}(A; \mathcal{L}(B; A \otimes B))$ (example 5), and from this last to $\mathcal{L}(B, A; A \otimes B)$ (example 5), and from this last to $\mathcal{L}(B \otimes A; A \otimes B)$ (example 6). Composing these and applying to the identity in $\mathcal{L}(A \otimes B; A \otimes B)$, we obtain, τ. Similarly, "the tensor product is associative".

Example Denote by τ the element of $\mathcal{L}(\mathcal{L}(A; B) \otimes C; \mathcal{L}(A; B \otimes C))$ given in the last example on page 41. We show that this τ is natural. Applying example 5 to the natural element of $\mathcal{L}(B, C; B \otimes C)$, we obtain a natural element of $\mathcal{L}(B; \mathcal{L}(C; B \otimes C))$. Applying this with the natural element of $\mathcal{L}(A, \mathcal{L}(A; B); B)$, we obtain a natural element of $\mathcal{L}(A, \mathcal{L}(A; B); \mathcal{L}(C; B \otimes C))$. But this last tensor space is naturally isomorphic with $\mathcal{L}(A, \mathcal{L}(A; B), C; B \otimes C)$, which is naturally isomorphic with $\mathcal{L}(\mathcal{L}(A; B), C; \mathcal{L}(A; B \otimes C))$, which is naturally isomorphic with $\mathcal{L}(\mathcal{L}(A; B) \otimes C; \mathcal{L}(A; B \otimes C))$. Applying these isomorphisms successively to the element of the previous sentence, we obtain τ.

It is easy to think of many other, similar, example, with similar proofs.

Finally, we remark that this unsettled state of the issue of what structure there is on the tensor spaces in no way directly affects what we shall do hereafter. We shall not, for example, attempt to use in proofs things which have not been proven. The one effect that we shall see is that we will not, occasionally, be able to state our conclusion in the pretty and general way that one might wish.

16. Natural Tensors: Continued

This section is just an appendix to the previous one. We wish to discuss two examples involving tensor spaces and natural tensors. In the first, we suggest a possible answer to the question of which tensor spaces are naturally isomorphic with which others. In the second, we indicate what happens to all the tensor spaces and natural tensors in the finite-dimensional case.

Fix a Banach space E. A tensor space A over E will be said to be in *canonical form* if it satisfies the following conditions: i) No "R" appears as a factor in a tensor product, or as an entry on the left in an "$\mathcal{L}$", ii) No entry on the left in an "$\mathcal{L}$" is the tensor product, of two other tensor spaces, and iii) No entry on the right in an "$\mathcal{L}$" is of the form $\mathcal{L}(B_i, \ldots, B_n; C)$. For example, none of the tensor spaces $\mathcal{L}(E; E \otimes R)$, $\mathcal{L}(R; E \otimes E)$, $\mathcal{L}(E \otimes E; R)$, or $\mathcal{L}(E; \mathcal{L}(E; R))$ are in canonical form, while $\mathcal{L}(E, E; R)$, $\mathcal{L}(\mathcal{L}(E; E); R)$, and $\mathcal{L}(E; E \otimes E)$ are all in canonical form.

We now claim that every tensor space is naturally isomorphic to a tensor space in canonical form. The proof is quite easy. Given tensor space A, one first uses the natural isomorphism between $R \otimes B$ and B to eliminate R's in tensor products. Then, using the natural isomorphism between $\mathcal{L}(A_1, \ldots, A_n, R; B)$ and $\mathcal{L}(A_1, \ldots, A_n; B)$ one eliminates R's on the left in $\mathcal{L}$'s. Then, using the isomorphism between $\mathcal{L}(A_1, \ldots, A_n, B \otimes C; D)$ and $\mathcal{L}(A_1, \ldots, A_n, B, C; D)$, one eliminates tensor products on the left in $\mathcal{L}$'s. Finally, using example 5, one eliminates $\mathcal{L}$'s on the right in $\mathcal{L}$'s. In this way, one obtains a sequence of natural isomorphisms which carry one from A to a tensor space in canonical form. Their composition is therefore the desired natural isomorphism.

This notion of a canonical form would not be very useful if we did not wish to claim some sort of converse to the result just proved. Specifically, we have:
Conjecture. Two tensor spaces in canonical form are naturally isomorphic if and only if they differ only in i) order of entries on the left in $\mathcal{L}$'s, and ii) order of, and parenthesis about, factors in tensor products.

The "if" part is of course obvious.

It would be nice to have a proof of this conjecture. Then, not only would every tensor space be naturally isomorphic to one in canonical form, but, furthermore, a given tensor space would be naturally isomorphic to only a single tensor space in canonical form (up to the ambiguity in the conjecture) (for, were A naturally isomorphic to two in canonical form, these two would be naturally isomorphic to each other). Thus, one would have a unique representative of each tensor space. One could then, for example, worry only about natural tensors in tensor spaces in canonical form. Although a proof of this conjecture from scratch appears not to be easy, there is, I should think, a reasonable chance that one could demonstrate that it is a corollary of the conjecture in the previous section (since that earlier conjecture claims to give all the natural tensors.).

In a similar way, one could try to formulate conjectures about what natural tensors there are, what natural subspaces, etc.

We consider, finally, the special case of all this when the Banach space E is finite-dimensional. In this case, everything is known, and everything is relatively easy: It is just what one learns in linear algebra. We shall here attempt only to indicate how the tensor spaces reduce to a simple form. We begin with the following observations: If A and B are finite-dimensional, then so is $A \otimes B$; If $A_1, \ldots, A_n$ and B are finite-dimensional, then so is $\mathcal{L}(A_1, \ldots, A_n; B)$. Thus, all the tensor spaces, when E is finite-dimensional, are finite-dimensional.

We first associate, with each tensor space, a pair of non-negative integers, according to the following rules. With E, associate $(1, 0)$; with R, associate $(0, 0)$; if with A there is associated (p, p'), and with B (q, q'), associate with $A \otimes B$ $(p+q, p'+q')$; if with $A_1, \ldots, A_n$

there are associated $(p_1, p'_1), \ldots, (p_n, p'_n)$, and with B there is associated (q, q'), then associate with $\mathcal{L}(A_1, \ldots, A_n; B)$ $(q + p'_1 + \ldots + p'_n, q' + p_1 + \ldots, +p_n)$, Using these rules, proceeding inductively, we associate a pair of integers with each tensor space. The pair of integers associated with tensor space A is called the *rank* of A. For example, the rank of the tensor space given in the middle of page 43 is $(4, 4)$.

The basic statement in the finite-dimensional case is this: Two tensor spaces over finite-dimensional E are naturally isomorphic if and only if they have the same rank. We shall actually prove slightly more than this. First, denote by E' the tensor space $\mathcal{L}(E; R)$. This E' is called the *dual* of Banach space E, and has rank $(0, 1)$. Given any pair (p, q) of non-negative integers, the tensor space $\mathcal{L}(E, \ldots, E, E', \ldots, E'; R)$, with pE''s and qE's, has rank (pq). Thus, in this way we obtain a particular tensor space of each rank. We shall show that, in the finite-dimensional case, any tensor space A of rank (p, q) is naturally isomorphic to the one just displayed.

We first claim the following: For A, B, and C any finite-dimensional tensor spaces, there is a natural isomorphism from $\mathcal{L}(A; B \otimes C)$ to $\mathcal{L}(\mathcal{L}(B; A); C)$. Proof: Choose bases for A, B, and C, say $x_1, \ldots, x_m, y_1, \ldots y_n$, and $z_1, \ldots, z_s$, respectively. Then an element of $B \otimes C$ is defined by an n s matrix a_{jk} $(j = 1, \ldots, n; k = 1, \ldots, s)$. Hence, an element of $\mathcal{L}(A; B \otimes C)$ is defined by a matrix a_{ijk} $(i = 1, \ldots, m; j = 1, \ldots, n; k = 1, \ldots, s)$, namely, such a matrix gives the element of $\mathcal{L}(A; B \otimes C)$ which sends $x = \Sigma_i b_i x_i$ in A to $\Sigma_i b_i a_{ijk}$ in $B \otimes C$. Similarly, an element of $\mathcal{L}(B; A)$ is represented by a matrix d_{ij} (i.e., sending $y = \Sigma_j b_j y_j$ in B to $\Sigma_{ij} b_j d_{ij} x_i$ in A). Hence, an element of $\mathcal{L}(\mathcal{L}(B; A); C)$ is represented by a matrics d_{ijk}. Thus, each of the tensor spaces $\mathcal{L}(A; B \otimes C)$ and $\mathcal{L}(\mathcal{L}(B; A); C)$ is represented by matrices, where these matrices are of the same kind. We thus obtain an isomorphism between these tensor spaces by comparing the matrix representations of their elements with respect to our bases for A, B and C. One easily checks that this isomorphism is independent of bases, and that it is natural. [It is interesting to note that, in the infinite-dimensional case, neither of $\mathcal{L}(A; B \otimes C)$ and $\mathcal{L}(\mathcal{L}(B; A); C)$ is even a subspaces of the other.]

Thus, in the finite-dimensional case, we have not only all the natural isomorphisms of the previous section, but also this additional one: $\mathcal{L}(A; B \otimes C)$ is naturally isomorphic with $\mathcal{L}(\mathcal{L}(B; A); C)$. From this, we obtain the following additional isomorphisms.

1. The tensor space $B \otimes C$ is naturally isomorphic with $\mathcal{L}(B'; C)$ (where prime denotes dual). Proof: Setting $A = R$ in the isomorphism above, we have that $\mathcal{L}(R; B \otimes C)$ is naturally isomorphic with $\mathcal{L}(\mathcal{L}(B; R); C)$. But $\mathcal{L}(R; B \otimes C)$ is naturally isomorphic with $B \otimes C$, while $\mathcal{L}(B; R)$ is just B'.

2. The tensor space B is naturally isomorphic with B'' (the dual of its dual). Setting $A = R$ and $C = R$ in the isomorphism above, we have that $\mathcal{L}(R; B \otimes R)$ is naturally isomorphic with $\mathcal{L}(\mathcal{L}(B; R); R)$. But the former is naturally isomorphic with B, and the latter is B''.

3. The tensor space $\mathcal{L}(A; B)$ is naturally isomorphic with $\mathcal{L}(\mathcal{L}(B; A); R)$. Proof: Set $C = R$ in the isomorphism above.

4. The tensor space $\mathcal{L}(A; B)$ is naturally isomorphic with $\mathcal{L}(A, B'; R)$. Proof: Substituting the result of (1) above in our basic isomorphism, we have that $\mathcal{L}(A; \mathcal{L}(B'; C))$ is naturally isomorphic with $\mathcal{L}(\mathcal{L}(B; A); C)$. But the first is naturally isomorphic with $\mathcal{L}(A, B'; C)$. Now set $C = R$. Then we have that $\mathcal{L}(\mathcal{L}(B; A); R)$ is naturally isomorphic with $\mathcal{L}(A, B'; R)$. By result (3) above, therefore, we have that $\mathcal{L}(A; B)$ is naturally isomorphic with $\mathcal{L}(A, B'; R)$.

Our claim, that every tensor space is the finite-dimensional case is naturally isomorphic to one of the form given at the bottom of the previous page, is now immediate. Given a tensor space, we first eliminate all tensor products by result (1) above. All $\mathcal{L}$'s which appear on the left in $\mathcal{L}$'s are than eliminated by (3). Next, anything except "R" appearing on the right in an "$\mathcal{L}$" is eliminated by (4). Finally, multiple duals are eliminated by (2). After all these eliminations, the only possible form is that given at the bottom of page 51. Furthermore, all of these "eliminations" preserve rank, as one check directly from (1)

– (4). We conclude, therefore, that every tensor space, in finite dimensions, is naturally isomorphic to the one of the form $\mathcal{L}(E, \dots, E, E', \dots, E'; R)$ of the same rank.

One could now continue in this way, finding all the natural tensors in the finite-dimensional case, etc. However, since there are no new ideas, since this amounts essentially only to restating what one knows from linear algebra in a slightly different terminology, and since we are not here particularly interested in finite dimensions, we go no further.

17. Tensor Fields

The examples (Sections 11 and 12) and the algebra (Sections 13, 14, 15, and 16) out of the way, we now turn to tensor fields on manifolds.

Let M be C^p ($p \geq 1$) manifold based on Banach space E. Fix a point p of M, and a tensor space A over E. Consider now pairs $(\alpha; U, \psi)$, where α is an element of the tensor space A, and U, ψ is an admissible chart on M, with p in U. Given two such, we write $(\alpha; U, \psi) \approx (\alpha'; U', \psi')$ if $\alpha' = \iota(\alpha)$, where ι is the isomorphism on tensor space A which arises (Section 15) from the isomorphism $\iota = D(\psi' \cdot \psi^{-1})(\psi(p))$ on E. This "$\approx$" is an equivalence relation. [The proof consists of a word-for-word repetition of the proof for tangent vectors on page 34, together with the observation that compositions and inverses of isomorphisms on E yield, when extended to the tensor space A, the compositions and inverses of the corresponding extension.] An equivalence class will be called an *$A-$tensor at p*. [Note that an $A-$tensor is just an element of the tensor space A, i.e., an element of a certain Banach space constructed from E using multilinear mappings and tensor products, while an $A-$tensor at p is an equivalence class of pairs, the first entry of each of which is an $A-$tensor. Sometimes, when we wish to emphasize the distinction, we shall call an element of the tensor space A a *free $A-$tensor*.]

We next note the following fact. Given any $A-$tensor ξ at p, and any admissible chart U, ψ with p in U, there is one and only one free $A-$tensor α with $(\alpha; U, \psi)$ in the equivalence class ξ. Indeed, letting $(\alpha; U', \psi')$ be any representative of the equivalence class ξ, the desired unique α is given by $\alpha = \iota^{-1}(\alpha')$, where the isomorphism ι on A comes from $D(\psi' \cdot \psi^{-1})(\psi(p))$ on E. This free $A-$tensor α will be called the *component* of ξ with respect to the chart U, ψ. [Motivation for the terminology: In the finite-dimensional case, one normally chooses a basis for the tensor space A, and thereby expressed the free $A-$tensor α in terms of n real numbers. Thus, an $A-$tensor at p defines, once a chart is given, n real numbers, numbers which are normally called the components of ξ. We do not choose bases, and hence replace these components by a single α in A.]

Example. Let τ be any natural tensor in tensor space A. Consider the collection of all pairs, $(\tau; U, \psi)$ whose first entry is this natural tensor. Since τ is $\iota-$invariant, any two such pairs are equivalent. Thus, we obtain an equivalence class, i.e., an $A-$tensor at p. This $A-$tensor at p has, of course, the property that its component with respect to any chart is just the free $A-$tensor τ. $A-$tensors at p so obtained will be called *natural $A-$tensors at p*.

What structure is there on the set of $A-$tensors at p? Picking any chart U, ψ with p in U, we obtain a one-to-one correspondence between the set of $A-$tensors at p and the tensor space A. By means of this correspondence, the entire structure of A – i.e., its structure as a Banach space – can be carried over to the set of $A-$tensors at p. We are interested only in that part of the structure which is chart-independent. Thus, exactly as with tangent vectors, the set of $A-$tensors at p has the structure of a Banachable space (real vector space, many equivalent norms all of which make it a Banach space). It turns out that, in fact, there is still more structure on the tensors at p. let τ be any natural operation, e.g., a natural element of the tensor space $\mathcal{L}(A_1 \ldots, A_n : B)$. Denote by σ the corresponding (example above) $\mathcal{L}(A_1, \ldots, A_n; B)-$tensor at p. Next, let $\kappa_1, \ldots, \kappa_n$ be $A_1-, \ldots, A_n-$tensors at p, respectively, Choose a chart, and let $\alpha_1, \ldots, \alpha_n$ be their respective components. Then $\tau(\alpha_1, \ldots, \alpha_n)$ is a free $B-$tensor. Now change the chart (U, ψ to U', ψ'. Then the components change to $\alpha_1 = \iota^{-1}(\alpha'_1), \ldots, \alpha_n = \iota^{-1}(\alpha'_n)$, while the components of σ does not change. We have $\iota^{-1}(\tau(\alpha_1, \ldots, \alpha_n)) = (\iota^{-1}(\tau))(\iota^{-1}(\alpha_1), \ldots, \iota^{-1}(\alpha_n)) = \tau(\alpha'_1, \ldots, \alpha'_n)$, where the first step is the action of ι on $\mathcal{L}(A_1, \ldots, A_n; B)$, and the second is definitions. But this equation is precisely the statement that $(\tau(\alpha_1, \ldots, \alpha_n) : U, \psi)$ is equivalent to $\tau(\alpha'_1, \ldots, \alpha'_n); U', \psi')$. Thus, we obtain a $B-$tensor at p. In short, natural operations are

extended in the obvious way from operations on free tensors to operations on tensors at p. We shall continue to use the term *natural operation* for these operations on tensors at p.

Example. "Apply", "compose", and "take the tensor product" are natural operations on tensors at p.

Our M continues to be a C^p ($p \geq 1$) manifold based on Banach space E. We next define fields. Let A be a tensor space over E. An $A-$field (on M) is a mapping which assigns, to each point p of M, an $A-$tensor at p. As usual, it is the smooth ones which are of most interest. Let ξ be an $A-$field on M, and let U, ψ be any admissible chart on M.

Then, for each point x of $\psi[U]$, $\psi^{-1}(x)$ is a point of M (in fact, of U), whence $\xi(\psi^{-1}(x))$ is an $A-$tensor at $\psi^{-1}(x)$, whence its components with respect to the chart U, ψ is an element of the Banach space A. Thus, we obtain a mapping from the open subset $\psi[U]$ of E to the Banach space A. The $A-$field ξ is said to be C^{p-1} if this mapping is C^{p-1} for every admissible chart U, ψ.

Example. Since ι is the identity on $\mathbb{R}$, we have that, for a, and a' reals, $(a : U\psi)\,(a'; U', \psi')$ if and only if $a = a'$. Thus, $\mathbb{R}-$tensors at p can be identified with real numbers. An $\mathbb{R}-$field on M is therefore a real-valued function on M. A C^{p-1} $R-$field on M is thus what we called a C^{p-1} scalar field in Section 11.

Example. A C^{p-1} $E-$field on M is what we called a C^{p-1} tangent vector field in Section 12.

Example. Let τ be any natural element of the tensor space A. Denote by σ the $A-$field on M such that, for each p in M, $\iota(p)$ is the $A-$tensor at p associated with this τ. This $A-$field is C^{p-1} (for the mapping above from $\psi[U]$ to A sends all of $\psi[U]$ to the single element τ of A, and this constant mapping is certainly C^{p-1}).

The (pointwise) sum of two C^{p-1} $A-$fields is a C^{p-1} $A-$field (since sums of C^{p-1} mappings of Banach spaces are C^{p-1}). We may also extend natural operations to the fields as follows. Let τ be a natural element of $\mathcal{L}(A_1 \ldots, A_n;\ B)$. Given $A_1-, \ldots, A_n-$fields $\kappa_1, \ldots, \kappa_{n'}$ we obtain, applying the construction of the middle of the previous page point wise, a $B-$field. If, furthermore, $\kappa_1 \ldots, \kappa_n$ are C^{p-1}, then so is this $B-$field (since a multilinear mapping of Banach spaces, applied to C^{p-1} mappings of Banach spaces, yields a C^{p-1} mapping). Thus, natural operations, applied point wise to C^{p-1} fields, yields C^{p-1} fields.

Example. We have "scalar multiplication", a natural element of $\mathcal{L}(\mathbb{R}, A; A)$. Hence, multiplication of $A-$fields by scalar fields yields $A-$fields. The last constructions on pages 31 and 38 are special cases.

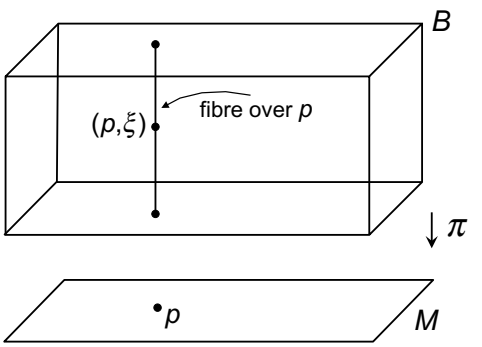
Example. The operations "applications". "composition", and "tensor product", applied to C^{p-1} filed, yields, yield C^{p-1} fields.

Example. From the first example and the observation that constant scalar fields (since they arise from natural tensors) are C^{p-1}, we have that set of $A-$fields form a vector space.

18. Tensor Bundles

Many notions in differential geometry have the feature that they possess two essentially equivalent formulations, one more analytic and algebraic, and the other more geometrical. One might even make a case that some of the interest in the subject arises from this interplay. We have already seen some examples of this phenomenon, e.g., in our discussion of tangent vectors. In the previous section, we treated what might be called the algebraic approach to tensor fields. We now give the geometric one.

Let M be a C^p ($p \geq 1$) manifold based on Banach space E, and fix a tensor space A over E. Denote by B the set of all pairs, (p, ξ), where p is a point of M, and ξ is an A-tensor at p. Denote by π the mapping from B to M which "ignores the second entry", i.e., with action $\pi(p, \xi) = p$. This B is called the *bundle space*, π the *projection mapping*, and M the *base space*. [In the figure, π is the mapping which "finds the point of M directly under the point of B".] For p a point of M the subset $\pi^{-1}[p]$ (i.e., the set of all elements of B of the form (p, ξ)) of B is called the *fibre* over p. This entire set-up (i.e., the bundle space, projection mapping, base space, fibres) is called the *A-bundle* of M.

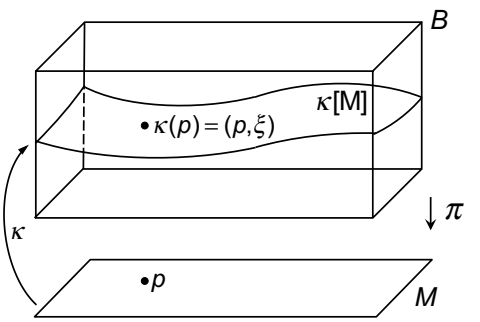


We have two goals: to find the structure and properties of the objects defined above, and to describe tensor fields in terms of these objects.

Consider first the bundle space B. We introduce some charts on this set. Let U, ψ be any chart on M. Set $\tilde{U} = \pi^{-1}[U]$, i.e., the union of the fibres over the points of U. Next, let $\tilde{\psi}$ be the mapping from the subset $\tilde{U}$ of B to the Banach space $E \times A$ with the following action: $\tilde{\psi}(p, \xi) = (\psi(p), \alpha)$, where α is component, with respect to the chart U, ψ, of the A-tensor ξ at p. We claim that this $\tilde{U}, \tilde{\psi}$ is an $E \times A$-chart on B. [Proof: Our $\tilde{\psi}$ is one-to-one, for, if $\tilde{\psi}(p, \xi) = \tilde{\psi}(p', \xi')$, then we must have $\psi(p) = \psi(p')$, and hence $p = p'$, and also $\alpha = \alpha'$, and hence $\xi = \xi'$. Further,since $\tilde{\psi}[\tilde{U}] = U \times A$ in $E \times A$, this $\tilde{\psi}[\tilde{U}]$ is open.] Now consider two of these charts on B, $\tilde{U}$, $\tilde{\psi}$ and $\tilde{U}'$,$\tilde{\psi}'$ (from, say, U, ψ and U', ψ' on M). We claim that these charts are C^{p-1}−compatible. [Proof: We have that $\tilde{\psi}[\tilde{U} \cap \tilde{U}'] = \psi[U \cap U'] \times A$, whence this is open in $E \times A$. The mapping $\tilde{\psi}' \cdot \tilde{\psi}^{-1}$ from $\tilde{\psi}[\tilde{U} \cap \tilde{U}']$ to $E \times A$ has action $\tilde{\psi}' \cdot \tilde{\psi}^{-1}(x, \alpha) = (\psi' \cdot \psi^{-1}(x), \kappa(x)(\alpha))$, where $\kappa(x)$ is the isomorphism on tensor space A arising from the isomorphism $D(\psi' \cdot \psi^{-1})(x)$ on E. Since composition, application, and insertion into the product are C^{p-1} operations, this mapping $\tilde{\psi}' \cdot \tilde{\psi}^{-1}$ is also C^{p-1}.] Thus, we now have a set B, together with a collection of C^{p-1}−compatible $E \times A$−charts on B. The first, second, and fourth conditions on page 25 are immediate from those conditions for M. We obtain, therefore, a C^{p-1} manifold B based on $E \times A$. [The idea is that "locally, B looks like a product of a small region of M with A".]

Thus, the bundle space B is a manifold. The base space is just M, and it starts out as a manifold. We have already investigated the structure of the fibers: The fibre over p is the set of A−tensors at p, which has the structure of a Banachable space. There remains, therefore, only the projection mapping π. Note that π now is a mapping of C^{p-1} manifolds (from B to M, which can be regarded as a C^{p-1} manifold). It would be natural to guess, therefore, that, this mapping will be C^{p-1}. It is. Proof: Let U, ψ be a chart on M, and $\tilde{U}, \tilde{\psi}$ the corresponding chart on B. Then $\psi \cdot \pi \cdot \tilde{\psi}^{-1}$ is the mapping from the open subset $\tilde{\psi}[\tilde{U}]$ of $E \times A$ to E with action $\psi \cdot \pi \cdot \tilde{\psi}^{-1}(x, \alpha) = x$. But this mapping, "projection onto the

first factor", is certainly C^{p-1}. We have shown that π, when "made a mapping of Banach spaces" via certain charts on B and M, yields a C^{p-1} mapping. But, since these "certain charts" cover B and M, the same holds for all charts (since all other charts are compatible with our "certain ones"). Thus, π is a C^{p-1} mapping of manifolds.



To summarize, then, the bundle space is a C^{p-1} manifold, the base space a C^p manifold, and the projection mapping C^{p-1}. The fibres are Banachable spaces.

This completes our first goal. We now describe tensor fields in this language. By a *cross section* of our $A-$bundle we mean a C^{p-1} mapping κ from M to B such that $\pi \cdot \kappa$ is the identity on M. That is to say, we require that, for each p in M, $\kappa(p)$ in B be of the form (p, ξ). Pictorially, one represents a cross-section by drawing $\kappa[M]$ in B as in the figure. We next note that a cross section κ of our $A-$bundle defines an $A-$field on M, for each p in M, the "ξ" of $\kappa(p) = (p, \xi)$ is an $A-$tensor at p. We claim: The $A-$field so obtained is C^{p-1}. Indeed, choosing chart U, ψ on M, and corresponding chart $\tilde{U}, \tilde{\psi}$ on B, we have since κ is C^{p-1}, that $\tilde{\psi} \cdot \kappa \cdot \psi^{-1}$ is a C^{p-1} mapping from E to $E \times A$. But the action of this mapping sends x in $\psi[U]$ to (x, α), where α is the component of ξ in the chart U, ψ. Hence, the mapping from $\psi[U]$ to A which "evaluates ξ and takes the component" is C^{p-1}. But this is precisely the statement that the $A-$field ξ is C^{p-1}.

We next claim the converse. Let ξ be a C^{p-1} $A-$field, and let κ be the mapping from M to B which action $\kappa(p) = (p, \xi(p))$. This κ, we claim, is a cross section. That $\pi \cdot \kappa$ is the identity is obvious, so we need only check $C^{p-1}-$ness. But, representing the mapping κ in terms of charts as in the previous paragraph, we have, since $\tilde{\psi} \cdot \kappa \cdot \psi^{-1}$ is C^{p-1} (this being what it means for the $A-$field ξ to be C^{p-1}, that κ is C^{p-1} (for this is what it means for a mapping of manifolds to be C^{p-1}).

We conclude: Cross section of the $A-$bundle are precisely the same things as (C^{p-1}) $A-$fields. In this sense, then, we "represent the fields geometrically". We shall see later that these bundles also permit one to draw pictures for various constructions, etc. involving tensor fields.

19. Lie Derivatives

We have now completed our discussion of three broad areas: calculus (Sects 2-8), manifolds (Sects 9-10), and tensor fields (Sects 11-18). The next broad area is that of derivatives (of tensor fields, on manifolds). That is to say, we now wish, for the first time, to make essential use of the symbols "C^p", "C^{p-1}", etc., that we have been carrying along. The basic problem is this: One is given, on a manifold, a certain tensor field, and one wishes to define a new tensor field which, can be interpreted as "the derivative" (in some sense, e.g., with respect to position on the manifold) of the original tensor field.

There is a naive way in which one might attempt to define "derivative" of a tensor field. It is of some interest, first of all, to see why it does not work. Let M be a C^p ($p \geq 2$) manifold based on Banach space E, let A be a tensor space over E, and let α be a C^{p-1} $A-$field. Fix a point p of M: We try to define "the derivative of α at p" as follows. Let U, ψ be any admissible chart on M, with p in U, and set $x = \psi(p)$. Then the component of our field with respect to this chart is a mapping $\hat{\alpha}$ from the open subset $\psi[U]$ of the Banach space E to the Banach space A. [That is to say, $\hat{\alpha}$ is the following. For q any point of U, $\hat{\alpha}(\psi(q))$, an element of Banach space A, is the component of the $A-$tensor $\alpha(q)$ at q with respect to the chart U, ψ.] The derivative of this mapping, $D\hat{\alpha}$, is thus a mapping from $\psi[U]$ to the Banach space $\mathcal{L}(E; A)$. Hence, $D\hat{\alpha}(x)$ is an element of $\mathcal{L}(E; A)$. [That is to say, we first "pull the field α over, using a chart, to a mapping from $\psi[U]$ to A". We then "take the derivative of this mapping with respect to the independent variable (point of $\psi[U]$)", and then, finally, we evaluate this derivative at $x = \psi(p)$, the image of p by ψ.] Consider now the pair, $(D\hat{\alpha}(x); U, \psi)$, consisting of a free $\mathcal{L}(E; A)-$tensor and a chart U, ψ. This pair certainly defines one of our equivalence classes, and hence defines a certain $\mathcal{L}(E; A)-$tensor at p. One would like to regard this $\mathcal{L}(E; A)-$tensor at p as "the derivative of the field α on M, evaluated at p". [Note that it is the right sort of object to be so regarded. "Locally near p, M looks like E." Hence, the "derivative of α with respect to position in M" should be a linear mapping from E (which represents "the space of possible directions in which one might move away from p in M") to A. That is, for ξ any tangent vector at p, and end denoting by β the $\mathcal{L}(E; A)-$tensor at p just obtained, then $\beta(\xi)$, an $A-$tensor at p, would be interpreted as "the derivative of the field α, in the $\xi-$direction, at p".]

There remains only one little check: chart-independence. If this check should work out properly, then the entire subject "derivatives of tensor fields" would be quite easy (consisting of this and the previous page), we could move on to a new topic, and differential geometry itself would have quite a different character than it has. Of course, it turns out that this check fails. The question, then, is this: Consider a new chart, U', ψ', with p in U'. Then, as above, we obtain a new point, $x' = \psi'(p)$, of $\psi'[U']$ (noting that this x' corresponds to the same p in M), a new mapping $\hat{\alpha}'$ from $\psi'[U']$ to A (noting that this $\hat{\alpha}'$ corresponds to the same $A-$field α), and a new pair, $(D\hat{\alpha}'(x'); U', \psi')$. Is it true or false that $(D\hat{\alpha}'(x'); U', \psi') \approx (D\hat{\alpha}(x); U, \psi)$, i.e., that we obtain the same $\mathcal{L}(E; A)-$tensor at p via U', ψ' as via U, ψ? To simplify this little calculation, set $A = E$. For q any point of $U \cap U'$, we have, setting $y' = \psi'(q)$ and $y = \psi(q)$, that $\hat{\alpha}'(y') = \tilde{\iota}(\hat{\alpha}(y))$, where $\tilde{\iota} = D(\psi' \cdot \psi^{-1})(y)$ (the formula for how the component of an $E-$tensor at q changes under change of chart). Setting $\iota = D(\psi' \cdot \psi^{-1})$ we may rewrite this formula as $\hat{\alpha}' \cdot (\psi' \cdot \psi^{-1})(y) = \iota(y)(\hat{\alpha}(y))$. Each side of this equation is a mapping from a certain open subset of $\psi[U]$ (that is where the variable y lives) to the Banach space $\mathcal{L}(E; E)$. Taking the derivative of this equation, using on the left the chain rule and on the right the Leibnitz rule for the derivative of the composition of two $y-$dependent mappings, we have that $D\hat{\alpha}'(\psi' \cdot \psi^{-1}(y)) \cdot D(\psi' \cdot \psi^{-1})(y) = D\iota(y)(\hat{\alpha}(y)) + \iota(y)(D\hat{\alpha}(y))$. Now set $y = x$ (so $\psi' \cdot \psi^{-1}(y) = x'$), and use the definition of ι for the second expression on the left to

obtain $D\hat{\alpha}'(x') \cdot \iota(x) = D\iota(x)(\hat{\alpha}(x)) + \iota(x)(D\hat{\alpha}(x))$. Finally, applying $\iota(x)^{-1}$ to both sides, we obtain our desired equation: $\iota(x)^{-1}D\hat{\alpha}'(x')\iota(x) = \iota(x)^{-1}D\iota(x)(\hat{\alpha}(x)) + D\hat{\alpha}(x)$. Now, this last formula is just true. What is it that we want to show? It is that $(D\hat{\alpha}'(x'); U', \psi') \approx (D\hat{\alpha}(x); U, \psi)$. Using the action of ι on $\mathcal{L}(E; E)$, this amount precisely to showing that $\iota(x)^{-1}D\hat{\alpha}'(x')\iota(x) = D\hat{\alpha}(x)$. Comparing what we have and what we want, we see that there is an extra term in the former $(\iota(x)^{-1}D\hat{\alpha}(x)(\alpha(x)))$ – a term which will not in general be zero. We conclude that our check fails. In general, we shall not obtain, by the prescription of the previous page, an $\mathcal{L}(E : A)-$ tensor at p which is independent of the chart used in that prescription.

All the mapping above tend to obscure what is basically a simple idea. The problem is that the "change in component under change in chart" mapping ι in general will depend on y. To pass from our $A-$field to our proposed $\mathcal{L}(E; A)-$field, we must take a derivative. Thus, although the component of α behaves properly (i.e., algebraically in ι) under chart-change, the "derivative of component" picks up an extra term involving the derivative of ι. This extra term prevents the "derivative of component" from behaving properly under chart-change, i.e., prevents "derivative of component" from representing the component of some other field on M.

The problem of taking derivatives of tensor fields amounts essentially to the problem of finding various ways to get rid of the "extra term" in the little calculation above. It turns out that there are at least three such ways, where the corresponding derivatives are called Lie derivatives, exterior derivatives, and derivative operators. Each of these types of derivative has its own advantages and disadvantages. In general terms, the advantage of each is that it "looks and acts like a derivative operation", and the disadvantage that some extra structure or restriction of the action has been employed to eliminate the extra term we found above. It is our proposal now to study these three kinds of derivatives, beginning with the Lie derivative.

It will be necessary, below, to make use of the following mapping. Let E be any Banach space, and A any tensor space over A. Denote by $\mathcal{L}_{\text{inv}}(E; E)$ the set of all invertible elements of $\mathcal{L}(E; E)$ (an open subset of the Banach space $\mathcal{L}(E; E)$, and hence a C^∞ manifold based on $\mathcal{L}(E; E)$). Next, let φ denote the following mapping from $\mathcal{L}_{\text{inv}}(E; E) \times A$ to A: For ι in $\mathcal{L}_{\text{inv}}(E; E)$ and α in A, $\varphi(\iota, \alpha)$ is the element of A obtained by first extending the isomorphism ι on E to A (Sect. 15), and then applying the resulting isomorphism from A to A to the element α of A. Thus, for example, we have $\varphi(\iota \cdot \iota', \alpha) = \varphi(\iota, \varphi(\iota', \alpha)), \varphi(\iota, \alpha + \alpha') = \varphi(\iota, \alpha) + \varphi(\iota, \alpha')$, and $\varphi(\iota, \alpha) = \alpha$ if α, is natural. This φ is a C^∞ mapping of manifolds. Hence, fixing α, the mapping φ_α from $\mathcal{L}_{\text{inv}}(E; E)$ to A with action $\varphi_\alpha(\iota) = \varphi(\iota, \alpha)$ is also C^∞. The derivative of this mapping, $D\varphi_\alpha$, is therefore a mapping from $\mathcal{L}_{\text{inv}}(E; E)$ to $\mathcal{L}(\mathcal{L}(E; E); A)$. Applying to the identity I (certainly an element of $\mathcal{L}_{\text{inv}}(E; E)$), we have $D\,\varphi_\alpha(I)$, an element of $\mathcal{L}(\mathcal{L}(E; E); A)$. Finally, let κ be the mapping from $\mathcal{L}(E; E) \times A)$ to A with action $\kappa(\xi, \alpha) = D\varphi_\alpha(I)(\xi)$. This κ is, clearly, bilinear, i.e., is an element of $\mathcal{L}(\mathcal{L}(E; E), A; A)$.

Example. Let $A = R$. Then φ has action $\varphi(\iota, r) = r$ (action of our isomorphisms on the reals), whence φ_r is the constant mapping (i.e., $\varphi_r(\iota) = r$ for every ι. Thus, $D\varphi_r = 0$, whence $\kappa = 0$.

Example. Let $E = E$. Then φ has action $\varphi(\iota, x) = \iota(x)$. Hence (since φ_x is a linear mapping), $D\varphi_x(\iota, \xi) = \xi(x)$, for ξ in $\mathcal{L}(E; E)$. Therefore, $\kappa(\xi, x) = \xi(x)$.

Example. Let $A = \mathcal{L}(E; E)$. Then φ has action $\varphi(\iota, \alpha) = \iota \cdot \alpha \cdot \iota^{-1}$, whence $D\varphi_\alpha(\iota)(\xi) = \xi \cdot \alpha \cdot \iota^{-1} - \iota \cdot \alpha \cdot \iota^{-1} \cdot \xi \cdot \iota^{-1}$. Therefore, $\kappa(\xi, \alpha) = \xi \cdot \alpha - \alpha \cdot \xi$.

Now fix, once and for all, a C^p ($p \geq 2$) manifold M based on Banach space E. The Lie derivative will be essentially a "generalized directional derivative" (generalized from action on scalar fields to all tensor fields on M). We shall need some direction in which to take the derivative. Hence, let ξ be a C^{p-1} tangent vector field ($E-$field) on M. Finally, let A be any tensor space over E, and let α be a C^{p-1} $A-$field on M. The idea is to define, from ξ and α, a new $A-$field on E.

Let U, ψ be a chart on M, and let η and β be the components of ξ and α, respectively

(so η is a C^{p-1} mapping from $\psi[U]$ to E, while β is a C^{p-1} mapping from $\psi[U]$ to A). Consider now the mapping ω from $\psi[U]$ to A with action $\omega(x) = D\beta(x)(\eta(x)) - \kappa(D\eta(x), \beta(x))$ (noting that all this is well-defined: $D\beta(x)$ is an element of $\mathcal{L}(E; A)$, whence $D\beta(x)$ $(\eta(x))$ is in A : $D\eta(x)$ is an element of $\mathcal{L}(E; E)$ and $\beta(x)$ an element of A, whence $\kappa(D\eta(x), \beta(x))$ is an element of A). Thus, we have so far written down a mapping ω from $\psi[U]$ to A. This ω is our candidate for the component of a certain $A-$field on M, namely, the component with respect to the chart U, ψ. To verify that this ω actually leads to an $A-$field, we must check to see how ω changes when the chart is changed. To this end, let $U'.\psi'$ be a different chart. Then, denoting by η' and β' the components of ξ and α, respectively, with respect to this chart, we have that $\beta'(x') = \varphi(\iota(x), \beta(x))$ and $\eta'(x') = \iota(x)(\eta(x))$, where $x' = \psi' \cdot \psi^{-1}(x)$ (the statement that the point x' of $\psi'[U']$ defines the same point of M as the point x of $\psi[U]$) and $\iota = D(\psi' \cdot \psi^{-1})$. These two formulae, then, are just those for the behaviour of the component under chart-change. Rewriting the first equation in the form $\beta' \cdot (\psi' \cdot \psi^{-1})(x) = \varphi_{\beta(x)}(\iota(x))$, taking the derivative (with respect to x), and applying to an arbitrary element v of E, we obtain $D\beta'(x')\iota(x)(v) = \varphi_D\beta(x)(v)(\iota(x)) + \kappa((D\iota(x)(v))\iota^{-1}(x), \varphi(\iota(x)\beta(x))$ where we used the chain rule on the left, the formula for the derivative of the application of a bilinear mapping to $x-$dependent vectors on the right, and the definition of κ in the last term. Now set $v = \eta(x)$ in this equation, to obtain $D\beta'(x')(\eta'(x')) = \varphi(\iota(x), D\beta(x)(\eta(x))) + \kappa(D\iota(x)(\eta(x))\iota^{-1}(x)$, $\varphi(\iota(x), \beta(x)))$. Next, take the derivative of the formula for the component change of ξ to obtain $D\eta'(x')\iota(x)(v) = (D\iota(x)(v))(\eta(x)) + \iota(x)(D\eta(x)(v))$, where v is an arbitrary element of E. Using the fact that mixed partials commute on the first term on the right, we have that this term is $(D\iota(x)(\eta(x))(v)$. Inserting this into the expression above, and using the fact that v is arbitrary, we have $D\eta'(x')\iota(x) = D\iota(x)(\eta(x)) + \iota(x)(D\eta(x))$. That is to say, $D\iota(x)(\eta(x))\iota^{-1}(x) = D\eta'(x') - \iota(x)(D\eta(x))\iota^{-1}(x)$. Substituting this into the formula on line seventeen above, we obtain $D\beta'(x')(\eta'(x')) = \varphi(\iota(x), D\beta(x)(\eta(x))) +$ $\kappa(D\eta'(x'), \varphi(\iota(x), \beta(x)) - \kappa(\iota(x)D\eta(x)\iota^{-1}(x), \varphi(\iota(x), \beta(x))$. But the second term on the right is just $\kappa(D\eta'(x'), \beta'(x')))$. Substituting, we obtain finally, $D\beta'(x')(\eta'(x')) -$ $\kappa(D\eta'(x'), \beta'(x')) = \varphi(\iota(x), D\beta(x)(\eta(x)) - \kappa(D\eta(x), \beta(x))$. But this formula is precisely the statement that $\omega'(x') = \varphi(\iota(x), \omega(x))$, i.e., the statement that, for each x, $(\omega'(x'); U', \psi') \approx (\omega(x); U, \psi)$. We conclude, therefore, that our ω indeed has the proper behaviour under changes in chart.

We now define the *Lie derivative*, $\mathcal{L}_\xi\alpha$, of the $A-$field α in the $\xi-$direction by the formula above. That is to say, letting η, β, and ω be the components of ξ, α, and $\mathcal{L}_\xi$ respectively, with respect to a chart, we set $\omega(x) = D\beta(x)(\eta(x)) - \kappa(D\eta(x), \beta(x))$.

The calculation above is rather messy because of its generality, and because there are so many mappings around. It is easier to see what is going on by looking at examples.

Example. Set $A = \mathbb{R}$. Thus, we wish to take the Lie derivative of scalar field α in the $\xi-$direction. In this case, β is a mapping from $\psi[U]$ to $\mathbb{R}$, and κ is zero. Hence, the formula above becomes $\omega(x) = D\beta(x)(\eta(x))$. But the right side will be recognized as the formula for the directional derivative of α in the direction of the tangent vector $\xi(p)$ at $p = \psi^{-1}(x)$. Thus, the Lie derivative of a scalar field is its directional derivative.

Example. Set $A = E$. We wish to take the Lie derivative of the tangent vector field α in the $\xi-$direction. Now κ has action $\kappa(\zeta, \nu) = \zeta(\nu)$, whence our formula above becomes $\omega(x) = D\beta(x)(\eta(x)) - D\eta(x)(\beta(x))$. This ω,then, is the component of $\mathcal{L}_\xi\alpha$. Note in particular from this formula that $\mathcal{L}_\xi\alpha = -\mathcal{L}_\alpha\xi$.

Example. Set $A = \mathcal{L}(E; E)$. Then κ has action $\kappa(\zeta, \mu) = \zeta \cdot \mu - \mu \cdot \zeta$. Hence, the formula for the component of $\mathcal{L}_\xi\alpha$ is $\omega(x) = D\beta(x)(\eta(x)) - D\eta(x) \cdot \beta(x) + \beta(x) \cdot D\eta(x)$.

It is a good exercise to verify explicitly that the $\mathcal{L}_\xi\alpha$ of each example above has the proper behavior under changes of chart.

We now have a thing, the Lie derivative, with a rather complicated definition. We deal with this situation in the usual way: We attempt to find a list, of reasonable length, of properties of the Lie derivative, with the goal of using in practice the properties on the

list rather than the original definition. Such a list follows.

Property 1. The Lie derivative of a C^{p-1} $A-$field in the $\xi-$direction (where ξ is a C^{p-1} $E-$field) is a C^{p-2} $A-$field. More generally, if the $A-$field is C^q $(1 \leq q \leq (p-1)$, or, for $A = \mathbb{R}$, $1 \leq q \leq p)$, and ξ is $C^{q'}$ $(1 \leq q' \leq (p-1))$, then $\mathcal{L}_\xi \alpha$ is $C^{q''}$, where q'' is the minimum of $q-1$ and $q'-1$. [This fact is immediate from the formula, and the fact that composition, application, etc. are C^∞ operations.]

Property 2. For α an $R-$field, $\mathcal{L}_\xi \alpha$ is the directional derivative of α in the $\xi-$direction. (Example, page 59.)

Property 3. For α an $R-$field, and ξ and τ $E-$fields, we have $\mathcal{L}_\xi(\mathcal{L}_\tau \alpha) - \mathcal{L}_\tau(\mathcal{L}_\xi \alpha) = \mathcal{L}_{\mathcal{L}_\xi \tau}\alpha$. Proof: Choose a chart, and let the components of ξ, τ and α be η, σ, and β, respectively. Then the component of $\mathcal{L}_\tau \alpha$ sends x to $\omega(x) = D\beta(x)(\sigma(x))$, whence the component of $\mathcal{L}_\xi(\mathcal{L}_\tau \alpha)$ sends x to $D\omega(x)(\eta(x)) = DD\beta(x)(\sigma(x), \eta(x)) + D\beta(x)[D\sigma(x)(\eta(x))]$. Reversing the roles of ξ and τ, the component of $\mathcal{L}_\xi \mathcal{L}_\tau \alpha - \mathcal{L}_\tau \mathcal{L}_\xi \alpha$ sends x to $DD\beta(x)(\sigma(x), \eta(x)) + D\beta(x)$
$[D\sigma(x)(\eta(x))) - DD\beta(x)(\eta(x), \sigma(x)) - D\beta(x)[D\eta(x)(\sigma(x))]$. But the first and third terms cancel, since mixed partials are symmetric, and so we obtain $D\beta(x)[D\sigma(x)(\eta(x)) - D\eta(x)(\sigma(x))]$. But this is precisely the component of $\mathcal{L}_{\mathcal{L}_\xi \tau}\alpha$. In fact, it is also true that $\mathcal{L}_\xi(\mathcal{L}_\tau \alpha) - \mathcal{L}_\tau(\mathcal{L}_\xi \alpha) = \mathcal{L}_{\mathcal{L}_\xi \tau}\alpha$ for any $A-$field α. The proof is essentially the same as that above: Write down both sides in terms of a chart.

Property 4. For α and α' $A-$fields, $\mathcal{L}_\xi(\alpha + \alpha') = \mathcal{L}_\xi \alpha = \mathcal{L}_\xi \alpha'$. This is immediate from the defining formula and the fact that κ is bilinear.

Property 5. For τ a $\mathcal{L}(A; B)-$field, and α an $A-$field, we have $\mathcal{L}_\xi(\tau(\alpha)) = (\mathcal{L}_\xi \tau)(\alpha) + \tau(\mathcal{L}_\xi \alpha)$. Proof: For $\hat{\tau}$ in $\mathcal{L}(A; B)$ and $\hat{\alpha}$ in A we have, taking the derivative of the action of ι on $\mathcal{L}(A; B)$, that $\kappa(\zeta, \hat{\tau}(\hat{\alpha}) = \kappa(\zeta\hat{\tau})(\hat{\alpha}) + \hat{\tau}(\kappa(\zeta, \hat{\alpha}))$ for every ζ in $\mathcal{L}(E; E)$. Denote by $\hat{\tau}$ and $\hat{\alpha}$ the components of τ and α, respectively. Then we have that the action of the component of $\mathcal{L}_\xi(\tau(\alpha))$ is $\omega(x) = D(\hat{\tau}(\hat{\alpha}))(x)\eta(x) - \kappa(D\eta(x), \hat{\tau}(x)(\hat{\alpha}(x)))$. Expand the first term on the right using the Leibnitz rule (for derivatives of mappings of Banach spaces), and use the fact above in the second. The result is the equation claimed, written in terms of our chart. By a similar argument one obtains: For τ a $\mathcal{L}(A_1, \ldots, A_n; B)-$field, and $\alpha_1, \ldots, \alpha_n A_1 -, \ldots, A_n-$fields, $\mathcal{L}_\xi(\tau(\alpha_1, \ldots, \alpha_n)) = (\mathcal{L}_\xi \tau)(\alpha_1, \ldots, \alpha_n) + \tau(\mathcal{L}_\xi \alpha_1, \alpha_2, \ldots, \alpha_n) + \ldots + \tau(\alpha_1, \ldots, \mathcal{L}_\xi \alpha_n)$. Thus, property 5 is the Leibnitz rule for Lie derivatives.

Property 6. For α any natural $A-$field, $\mathcal{L}_\xi \alpha = 0$. Proof: For α natural, and $\hat{\alpha}$ the component of α, we have that the mapping $\hat{\alpha}$ is constant (for $\hat{\alpha}(x)$ is this natural tensor for each x). Furthermore, $\kappa(\zeta, \hat{\alpha}(x)) = 0$ for any ζ in $\mathcal{L}(E; E)$. Hence, for the component of $\mathcal{L}_\xi \alpha$, we have $D\hat{\alpha}(x)(\hat{\xi}(x)) - \kappa(D\hat{\xi}(x), \hat{\alpha}(x)) = 0$ (the first term vanishing since it is a derivative of a constant map). Thus, since the component of $\mathcal{L}_\xi \alpha$ in every chart vanishes, we have $\mathcal{L}_\xi \alpha = 0$.

Property 7. The Leibnitz rule is satisfied for any natural tensor operation. That is, for τ natural in $\mathcal{L}(A_1, \ldots, A_n; B)$, and $\alpha_1, \ldots, \alpha_n A_1 -, \ldots, A_n-$tensors, $\mathcal{L}_\xi(\tau(\alpha_1, \ldots, \alpha_n)) = \tau(\mathcal{L}_\xi \alpha_1, \ldots, \alpha_n) + \ldots + \tau(\alpha_1, \ldots, \mathcal{L}_\xi \alpha_n)$. Proof: Immediate from properties 5 and 6.

Properties 4 and 7 summarize the dependence of "$\mathcal{L}_\zeta \alpha$" on α. We next consider its dependence on ξ. To this end, first note that, for any tensor space A, the element κ of $\mathcal{L}(\mathcal{L}(E; E), A; A)$ is natural. Hence we obtain, on our manifold, a corresponding natural $\mathcal{L}(\mathcal{L}(E; E), A; A)-$field, which we also denote κ.

Property 8. For α and $A-$field, and ξ and ξ' $E-$fields, we have $\mathcal{L}_{\xi+\xi'} \alpha = \mathcal{L}_\xi \alpha + \mathcal{L}_{\xi'} \alpha$. Immediate.

Property 9. Let α be an $A-$field, ξ an $E-$field, and f an $\mathbb{R}-$field. Denote by ζ the $\mathcal{L}(E; E)-$field with action $\zeta(\nu) = \mathcal{L}_\nu f)\xi$ for every $E-$field ν. Then $\mathcal{L}_{(f\xi}\alpha = f(\mathcal{L}_\xi \alpha) + \kappa(\zeta, \alpha)$. Proof: The component of $\mathcal{L}_{(f\xi)}\alpha$ has action $\omega(x) = D\hat{\alpha}(x)(f(x)\hat{\xi}(x)) - \kappa(D(f\hat{\xi})(x), \hat{\alpha}(x))$. Expanding the second term on the right using the Leibnitz rule for derivatives, this becomes $\omega(x) = D\hat{\alpha}(x)(f(x)\hat{\xi}(x)) - \kappa(f(x)D\hat{\xi}(x), \hat{\alpha}(x)) - \kappa(\zeta(x), \hat{\alpha}(x))$. Each of the first two terms on the right is now linear in $\xi(x)$, whence the $f(x)$, a number, can be pulled outside. The result is precisely the component form of the claimed equally.

This completes our list of properties.

Example. For α and $A-$field, and f and $\mathbb{R}-$fields, $\mathcal{L}_\xi(f\alpha) = (\mathcal{L}_\xi f)\alpha + f(\mathcal{L}_\xi\alpha)$. Property 7.

Example. For α an $A-$field and β a $B-$field, $\mathcal{L}_\xi(\alpha\otimes\beta) = (\mathcal{L}_\xi\alpha)\otimes\beta + \alpha\otimes(\mathcal{L}_\xi\beta)$. Property 7.

Example. For ξ and η $E-$fields, and f and $R-$field, $\mathcal{L}_{(f\xi)}\eta = f(\mathcal{L}_\xi\eta) + (\mathcal{L}_\eta f)\xi$. Property 9, or else properties 3 and 7.

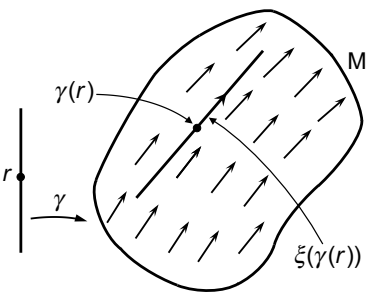
It is common in the finite-dimensional case to define the Lie derivative, not by the chart-formula that we have used here, but rather by some combination of the properties above. The idea is to show that there is one and only one "Lie derivative operation" satisfying certain properties, and then establish our chart-formula as a theorem. In more detail, one uses property 2 to get the action of the Lie derivative on scalar fields, then property 3 to get the action on vector fields, and finally various special cases of property 7 to get the action on other tensor fields. One could certainly carry out a similar program in the infinite-dimensional case but, unfortunately, a number of difficulties intervene to make this somewhat awkward. We here just mention what these difficulties are and how they might be surmounted. First, one's manifold may not admit a reasonable number of fields (as we have seen in an example, for scalar fields). One is thus forced, apparently, to introduce a "Lie derivative operation" locally, in small open regions, and then piece together these regions to get the operation over the entire manifold. Second, one must apparently impose all the differentiability assignments (property 1) on one's Lie derivative operation (including the exception for scalar fields). If one, for example, ignored the fact that scalar fields can be C^p (whereas others must be C^{p-1}), then one would not, by property 3, obtain a C^{p-2} vector field for the Lie derivative of a vector field. [The same problem arises in the finite-dimensional case. However, in finite dimensions one often works in C^∞, because it seems to turn out in practice that nothing is lost by such a restriction. Property 1 then simplifies somewhat. However, it is not so clear that "everything C^∞" is a reasonable condition in infinite-dimensions.] Third, one has to go to charts anyway, in the use of property 3. From this property, one will indeed find $(\mathcal{L}_\xi\eta)(f)$, once one knows Lie derivatives of scalar field. One must, however, then show that this is the directional derivative of f by some vector field (to be identify with $\mathcal{L}_\xi\eta$). This demonstration requires, apparently, charts. Fourth, there are apparently some problems with tensor products. We can of course define $\mathcal{L}_\xi(\alpha\otimes\beta)$ by the second example on the previous page. What, however, is the formula for $\mathcal{L}_\xi\tau$ when τ is an arbitrary $A\otimes B-$tensor? It is obvious to me that every such τ can be written in the form $\tau = \alpha_1\otimes\beta_1 + \alpha_2\otimes\beta_2 + \ldots$ (so that we could set $\mathcal{L}_\xi\tau = (\mathcal{L}_\xi\alpha_1)\otimes\beta_1 + \alpha_1\otimes(\mathcal{L}_\xi\beta_2) + \ldots$). Furthermore, even if this were true, we would still have to show that the second sum converges (in the tensor product) provided that the first one does. Smoothness would be a further problem. Finally, I am aware of no way, in this program, to show directly (i.e., without going back to charts) that the Lie derivative of every natural tensor vanishes. [This problem does not arise in the finite-dimensional case, since there one has the complete list of natural tensors, and simply checks them one at a time.]

It is our claim that all the various properties of the Lie derivative give expression to the idea that "the Lie derivative is a generalized directional derivative". The nice thing about the Lie derivative is that it can be applied to any tensor field, and it has a large number of properties. The disadvantage is that it requires some particular choice of a vector field, ξ, along which the Lie derivative is taken.

20. Integral Curves

The definition of the previous section may be called the algebraic-analytic approach to Lie derivatives. There is also a more geometrical approach to the same subject. It is our intention to discuss this alternative viewpoint. We require, as a prerequisite, the notion of an integral curve, and some properties of these. This general subject – integral curves – has numerous applications, extending far beyond just Lie derivatives, one of which we indicate briefly.

Fix, once and for all, a C^p ($p \geq 2$) manifolds M based on Banach space SE, and a C^{p-1} $E-$field ξ on M. An *integral curve* of ξ consists of an open interval, (a, b), of the reals, together with a C^p mapping γ from (a, b) (regarded as a manifold) to M, such that the following property is satisfied: For each number r in (a, b), the tangent vector to the curve γ at r is precisely $\xi(\gamma(r))$, i.e., the value of the field ξ at the point $\gamma(r)$ of M. Intuitively, an integral curve is "always moving along in M tangentially to ξ".



Example. Let (a, b),γ be an integral curve, and let s be any real number. Then $(a+s, b+s)$, $\tilde{\gamma}$ is also an integral curve,where $\tilde{\gamma}$ has action $\tilde{\gamma}(r) = \gamma(r - s)$. These two integral curves are said to be related by *reparameterization*.

Example. Let $M = E$. Fix any ξ_0 in E, and let ξ be that vector field on M whose component in this chart is ξ_0 for all x. Then, in terms of this chart, a typical integral curve is given by (a, b), γ, where $\gamma(r) = x_0 + r\xi_0$, where x_0 is any fixed vector in E.

It is our goal to decide whether or not integral curves exist, and how unique they are. It is convenient, for this discussion, to have available three definitions. For (a, b), γ an integral curve, with 0 in (a, b), the point $\gamma(0)$ of M is called the *initial point* of the curve. Clearly, e.g., a reparameterization serves merely to shift the initial point along the integral curve. Next, let (a, b), and (a', b'), γ' be two integral curves. The second is said to be an *extension* of the first if i) $a' \leq a$ and $b \leq b'$ (i.e., (a', b') includes (a, b)), and ii) $\gamma(r) = \gamma'(r)$ for r in (a, b) (i.e., wherever both curves are defined). Thus, an extension of an integral curve merely "makes the curve longer, on one or both ends". Finally, an integral curve, the only extension of which is the curve itself, is called *maximal*. [So, as one might expect, "is an extension of" is a partial ordering on integral curves.] Thus, for example, an integral curve is maximal if and only if every reparametrization is; the curve of the second example of page 62 is maximal provided $a = -\infty$ and $b = \infty$.

Clearly, integral curves are only going to be unique up to reparametrizations and extensions. The theorem on existence and uniqueness of integral curves is the simplest and strongest one could expect in light of this observation.

Theorem. Let M be a C^p ($p \geq 2$) manifold based on E, ξ a C^{p-1} $E-$field, and p a point of M. Then there exists one and only one maximal integral curve γ of ξ with initial point p.

Proof: We first make the following observation. Let U, ψ, be a chart on M, let $O = \psi[U]$, let $\hat{\xi}$ (a mapping from O to E) be the component of ξ, and let $\hat{\gamma} = \psi \cdot \gamma$ (a curve on O). Then the statement that γ is an integral curve becomes, in terms of this chart, the equation $D\hat{\gamma} = \hat{\xi} \cdot \hat{\gamma}$.

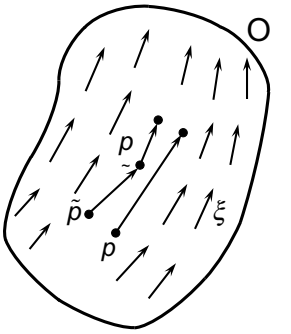
Uniqueness. Let O and $\hat{\xi}$ be as above, and let c and d be numbers such that $c = \text{lub}\, |\hat{\xi}(x)|$ and $d = \text{lub}\, |\mathrm{D}\hat{\xi}(\mathrm{x})|$. Let (a, b) be an open interval with $a < 0 < b$, and let $\hat{\gamma}$ and $\hat{\gamma}'$ be mappings from (a, b) to O satisfying $D\hat{\gamma} = \hat{\xi} \cdot \hat{\gamma}$ $D\hat{\gamma}' = \hat{\xi} \cdot \hat{\gamma}'$, and $\hat{\gamma}(\beta) = \hat{\gamma}'(0) = x_0$. Finally, let τ have action $\tau(x) = \hat{\gamma}(x) - \hat{\gamma}'(x)$. We have $D\tau(r) = D\hat{\gamma}(r) - D\hat{\gamma}'(r) = hat\xi(\hat{\gamma}(r)) - \hat{\xi}(\hat{\gamma}'(r)) \leq \text{lub}\, |\mathrm{D}\hat{\xi}|\, |\hat{\gamma}(\mathrm{r}) - \hat{\gamma}'(\mathrm{r})| = \mathrm{d}\, |\tau(\mathrm{r})|$, were we used the mean value theorem in the

third step and definitions in the fourth. Hence, $|\tau(r)| = |\tau(r) - \tau(0)| \leq \operatorname{lub}_{0\leq r'\leq r} |D\tau(r')|\,|r| \leq d \operatorname{lub}_{0\leq r'\leq r} |\tau(r')|\,|r|$. Now choose positive r_0 such that $r_0 d \leq 1/2$. Then, taking the lub of the formula just obtained, we have $\operatorname{lub}_{0\leq r\leq r_o} |\tau(r)| \leq dr_0 \operatorname{lub} |\tau(\mathrm{r})| \leq 1/2 \operatorname{lub}_{0\leq \mathrm{r}\leq \mathrm{r}_0} |\tau(\mathrm{r})|$. But this is possible only if $\operatorname{lub} |\tau(\mathrm{r})| = 0$ i.e., if $\tau(r) = 0$, for every r in $[), r_0]$. That is, $\hat{\gamma}(r) = \hat{\gamma}'(r)$ for $0 \leq r \leq r_0$.

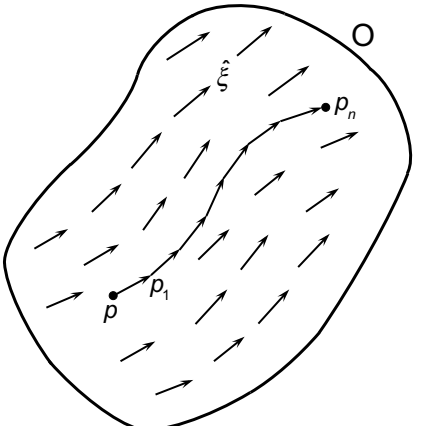
Now consider two maximal integral curves γ and γ' with $\gamma(0) = \gamma'(0) = p$, as in the theorem. Let s be the largest.

Now consider two maximal integral curves, (a, b), γ and (a', b'), γ', with $\gamma(0) = \gamma'(0) = p$, as in the theorem. Let s be the largest number such that $s \leq b$, $s \leq b'$, and $\gamma(r) = \gamma'(r)$ whenever $0 \leq r < s$. We claim that the assumption $s < b$ leads to a contradiction. Indeed, we must have, under this assumption, $s < b'$, for otherwise the integral curve γ would lead to an extension of γ', contradicting maximality of γ'. Furthermore, by continuity, we must have $\gamma(s) = \gamma'(s)$. Choosing a chart U, ψ in M, with $\gamma(s) = \gamma'(s)$ in U, we have, by the result of the previous paragraph, that $\gamma(r) = \gamma'(r)$ also for $s \leq r \leq s + r_0$, for some positive r_0. But this violates the definition of s. This contradiction establishes that $s = b$, and, similarly $s = b'$. That is to say, we have $b = b'$ and $\gamma(r) = \gamma'(r)$ for $0 \leq r < b$. In a similar way, "working in the other direction", we have that $a = a'$, and $\gamma(r) = \gamma'(r)$ for $a < r \leq 0$. That is to say, our two maximal integral curves are identical.

Existence. Let O be an open subset of Banach space E, $\hat{\xi}$ a C^{p-1} mapping from O to E, and let c and d be numbers such that $c = \operatorname{lub} |\xi(\mathrm{x})|$ and $d = \operatorname{lub} |\mathrm{D}\xi(\mathrm{x})|$. Finally, let p and $\tilde{p}$ be two points of O, let ϵ be a sufficiently small positive number, and let α be a number with $0 \leq \alpha \leq 1$. Set $s = |p - \tilde{p}|$. We now associate, with each of p and $\tilde{p}$, a new point of O as follows. With p, associate the point $p + \epsilon\hat{\xi}(p)$. With $\tilde{p}$, associate the point obtained by first finding $\tilde{p} + \alpha\epsilon\hat{\xi}(\tilde{p})$, and then $\tilde{p} + \alpha\epsilon\hat{\xi}(\tilde{p}) + (1-\alpha)\epsilon\hat{\xi}(\tilde{p} + \alpha\epsilon\hat{\xi}(\tilde{p}))$. [That is, the first point is obtained by "going amount ϵ along the direction of $\hat{\xi}(p)$ from p"; the second point by "first going amount $\alpha\epsilon$ along the direction of $\hat{\xi}(\tilde{p})$ from $\tilde{p}$ to obtain $\underset{\sim}{p} = \tilde{p} + \alpha\epsilon\hat{\xi}(\tilde{p})$, and then going amount $(1 - \alpha)\epsilon$ along the direction of $\hat{\xi}(p)$ from $\underset{\sim}{p}$".] We wish to find the "distance" s' between the two points so associated. We have $s' = |p + \epsilon\hat{\xi}(p) - \tilde{p} - \alpha\epsilon\hat{\xi}(\tilde{p}) - (1-\alpha)\epsilon\hat{\xi}(\tilde{p} + \alpha\epsilon\hat{\xi}(\tilde{p}))| = |p - \tilde{p} + \epsilon\hat{\xi}(p) - \epsilon\hat{\xi}(\tilde{p}) + (1-\alpha)\epsilon\hat{\xi}(\tilde{p}) - (1-\alpha)\epsilon\hat{\xi}(\tilde{p} + \alpha\epsilon\hat{\xi}(\tilde{p}))| \leq s + \epsilon\,|\hat{\xi}(p) - \hat{\xi}(\tilde{p})| + |(1-\alpha)\epsilon\hat{\xi}(\tilde{p}) - (1-\alpha)\epsilon\hat{\xi}(\tilde{p} + \alpha\epsilon\hat{\xi}(\tilde{p}))| \leq s + \epsilon\, d\, s + (1-\alpha)\epsilon\,|\hat{\xi}(\tilde{p}) - \hat{\xi}(\tilde{p} + \alpha\epsilon\hat{\xi}(\tilde{p}))| \leq s(1 + \epsilon\, d) + (1-\alpha)\epsilon(d\,|\alpha\epsilon\hat{\xi}(\tilde{p})|) \leq s(1 + \epsilon\, d) + (1-\alpha)\epsilon\, d\,\alpha\epsilon\, c \leq s(1 + \epsilon\, d) + \epsilon^2 cd/4$, where the first step is definition of s', the second results from rearranging terms, the third is the triangle inequality, the fourth results from using the mean value theorem on the second term, the fifth results from using the mean value theorem on the third term, the sixth uses the definition of c, and the seventh uses the fact that $\alpha(1 - \alpha) \leq 1/4$. We shall use this formula, $s' \leq s(1 + \epsilon\, d) + \epsilon^2 cd/4$, several times in what follows.



Next, fix a point p of O and a sufficiently small positive number $\underline{r}$. Let $r_0, \ldots, r_n$ be numbers with $0 = r_0 < r_1 < \ldots < r_n = \underline{r}$, and set $\epsilon = \max |\mathrm{r_{i+1}} - \mathrm{r_i}|$. We construct, from this set-up, a point of O as follows: First find $p_1 = p + r_1\hat{\xi}(p)$, then $p_2 = p_1 + (r_2 - r_1)\hat{\xi}(p_1)$, then, $p_3 = p_2 + (r_3 - r_2)\hat{\xi}(p_2)$, etc. Continue, until one finds p_n. Now let there be given numbers $\alpha_1, \ldots, \alpha_n$, all between zero and one. Then we can find a second point of O by using the α's to subdivide as above. That is, set $\tilde{p}_1 =$



$p + \alpha_1 r_1 \hat{\xi}(p) + (1-\alpha_1) r_1 \hat{\xi}(p + \alpha_1 r_1 \hat{\xi}(p))$, then
$\tilde{p}_2 = \alpha_2 r_2 \hat{\xi}(\tilde{p}_1) + (1-\alpha_2)(r_2 \hat{\xi}(\tilde{p}_1 + \alpha_2 r_2 \hat{\xi}(\tilde{p}_1))$,
etc. Set $s_i = |p_i - \tilde{p}_i|$, and write μ for $1 + \epsilon d$
and ν for $\epsilon^2 cd/4$. Then, by our formula above,
$s_1 \leq \nu$. Applying the same formula for the second step, we have $s_2 \leq \mu s_1 + \nu \leq \nu(1+\mu)$. For the third, $s_3 \leq \mu s_2 + \nu \leq \nu(1+\mu+\mu^2)$. Continuing in this way, $s_n \leq \nu(1+\mu+\ldots+\mu^{n-1}) = \nu(\mu^n - 1)/(\mu - 1) = 1/4\epsilon c[(1+\epsilon d)^n - 1] \leq 1/4\epsilon c[c^{\epsilon n d} - 1]$, where the second and fourth steps use facts from elementary algebra, and the third is substitution for μ and ν.

Define the efficiency of a partition $r_0, \ldots, r_n$ as above to be the number $\underline{r}/\epsilon n$ (i.e., a measure of how "nearly equally spaced the intervals are", clearly always between 0 and 1). Restricting ourselves to partitions with efficiencies at least $1/2$, the formula of the previous paragraph becomes $s_n \leq 1/4\epsilon c[e^{2\underline{r}d} - 1]$. Now consider a sequence of such partitions, each having with efficiency at least $1/2$, each having an "n" twice that of its predecessor, and each being a refinement of its predecessor in the sense of the previous paragraph. Then, by this last formula, the sequence of endpoints obtained for each by our construction will be Cauchy (since this formula requires that the distance between two successive endpoints is bounded by some constant times the "ϵ" of the first partition), and hence will converge to some point of O. Repeating for different values r of $\underline{r}$ we obtain a mapping γ which assigns, to each sufficiently small positive r, a point $\gamma(r)$ of O.

We next claim that this mapping γ is C^1, and satisfies $D\gamma = \hat{\xi} \cdot \gamma$. Clearly, it suffices to check differentiability at O. We have $|\gamma(\underline{r}) - p - \underline{r}\hat{\xi}(p)| \leq \underline{r}c(e^{2\underline{r}d} - 1)$, since the left side asks for the distance between the endpoints for a very fine partition, and the partition with just one step (i.e., using repeatedly the formula of the previous paragraph). But, as $\underline{r}$ goes to zero, $(e^{2\underline{r}d} - 1)$ also goes to zero, whence the left side is tangent at O. In other words, γ is differentiable, and $D\gamma = \hat{\xi} \cdot \gamma$. Since in particular γ is continuous, the right side of this last equation is continuous, whence $D\gamma$ is continuous, whence γ is C^1. Since γ is C^1, the right side is C^1, whence $D\gamma$ is C^1, whence γ is C^2. We continue in this way until we reach the differentiability class of $\hat{\xi}$. That is ... since γ is C^{p-1}, the right side is C^{p-1}, whence γ is C^p.

Now fix any point of our manifold M. Choosing a chart including this point, and using the construction above, we obtain an integral curve with this point as initial point. Extend this curve maximally by Zorn's Lemma. The resulting curve can have no endpoints (for if it had one, we could choose a chart including this endpoint, and use our construction to further extend the curve). This is our maximal integral curve.

This completes the proof of our theorem.

Conceptually, the proof above is extremely simple. For uniqueness, one uses the mean value theorem twice to show that "the two curves cannot get much farther apart than they had been previously". For existence, one constructs "broken straight lines, each segment of which is in the direction of the field at its initial point", and then takes a limit to obtain integral curves. I find it a bit strange that such simple ideas turn out to be so complicated when written out in detail. In the more conventional proof of this theorem, one rewrites "is an integral curve" as an integral equation, regards one side of that equation as a mapping from curves to curves, shows that the set of curves forms a complete metric space and that this mapping is a contraction mapping, and uses the contraction lemma. This proof is essentially the same as the one above. We have incorporated the definitions of "integral" and "contraction mapping" into the proof without mentioning these terms, and have proved directly in context the properties needed of integrals and contraction mappings. The above proof, while somewhat longer than that with integrals, would perhaps be comparable in length if the latter were accompanied by the definition and properties of integrals. The above proof appeals somewhat more to my intuition.

We first give a simple example just to illustrate how one uses the theorem.

Example. Two integral curves on a manifold cannot cross (i.e., have just one point in common), for, were this the case, one could reparametrize them so that the "crossing point" is the initial point of each, witch would violate uniqueness.

In the finite-dimensional case, the theorem above might be called "the fundamental existence and uniqueness theorem for solutions of systems of ordinary differential equations". We give one example to illustrate the appropriateness of this description, leaving the statement of the general case as an (easy) exercise.

Example. Consider the following system of ordinary differential equations. We are interested in functions x, y and z of real variable t, satisfying the equations

$\ddot{x} = (1 - \dot{z}^2 + y^2)^{-1} + t^2$

$\dot{y} = -\dot{x}^2 + x^2 + \log(z + \dot{x}$

$\ddot{z} = \cos(2x\dot{z}t)$

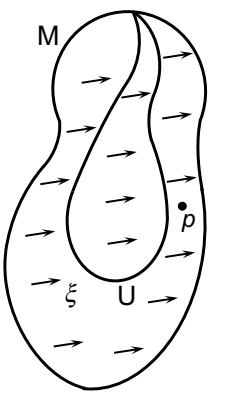
where "dot" means "d/dt". [Note that this is a "reasonable system", in the sense that the equations are solved for the highest derivative of each independent variable.] Denote by E the Banach space R^6, so a point of E is represented by six real numbers, $(x_1, \ldots, x_6)$. [Motivation: Think of $x_1 = x, x_2 = \dot{x}, x_3 = y, x_4 = z, x_5 = \dot{z}, x_6 = t$.] Next, let M be the (open sub-)manifold (of E) consisting of $(x_1, \ldots, x_6)$ satisfying $(1-(x_5)^2+(x_3)^2) \neq 0$ and $(x_4 + x_2) > 0$. [Motivation: We throw away precisely those points at which the right sides of the differential equations above are not well-behaved.] Next, let ξ be the mapping from M to E with action $\xi(x_1, \ldots, x_2) = x_2, (1-(x_5)^2+(x_3)^2)^{-1}+(x_6)^2, -(x_2)^2+(x_1)^2+(x_3)^2+\log(x_4+x_2), x_5, \cos(2x_1x_5x_6), 1)$. [Motivation: That the second, third, and fifth entries are what they are can be seen by looking at the right sides of our differential equation. For the remaining three entries, append the equations "$\dot{x} = \dot{x}$", "$\dot{z} = \dot{z}$", and "$\dot{t} = 1$" to our three differential equations.] This ξ is (with respect to the natural chart on M, the component of) a C^∞ vector field on M.

A curve on M maps some open interval (a, b) of the reals to M. Hence, for γ such a curve and r in (a, b), $\gamma(r)$ is a point of M, a point which can be represented by a six-tuple. That is to say, a curve can be described by six functions, $x_1, \ldots, x_6$, of one real variable r. The statement that this γ be an integral curve of our vector field is, of course, the statement that $dx_1/dr = x_2, dx_2/dr = (1 - (x_5)^2 + (x_3)^2)^{-1} + (x_6)^2, dx_3/dr = -(x_2)^2 + (x_1)^2 + (x_3)^2 + \log(x_4 + x_2), dx_4/dr = x_5, dx_5/dr = \cos(2x_1x_5x_6), dx_6/dr = 1$. We now claim that every such integral curve gives rise to a solution of our differential equation (namely, given an integral curve, replace r by t, and identify $x(t)$ with $x_1(r)$, $y(t)$, with $x_3(r), z(t)$ with $x_4(r)$, and t with $x_6(r)$. Then the first, fourth, and sixth of the equations above identify $x_2(r)$ with $\dot{x}(t), x_5(r)$ with $\dot{z}(t)$, and $x_6(r)$ with t. The other three equations are then precisely our original differential equations.) Conversely, any solution of our original system of differential equations gives rise to an integral curve, by the same identifications. Under our identification, specifying an initial value for an integral curve amounts to specifying initial values for x, $\dot{x}$, y, z, $\dot{z}$ and t. From our theorem, we conclude therefore: Given initial values of x, $\dot{x}$, y, z, $\dot{z}$ and t, there is one and only one solution of the system of differential equations on the previous page, maximally extended.

I regard it as an enormous conceptual simplification that "everything one always wanted to know about ordinary differential equations" is summarized so concisely by the theorem. One might expect to be able to do a similar thing for (e.g., hyperbolic) partial differential equations, where now the manifold M is the infinite-dimensional manifold of possible initial data for the equation. I feel that it would be of particular interest to investigate (e.g., beginning with some examples) such a program.

21. Geometry of Lie Derivatives

We now provide the promised geometrical interpretation of Lie derivatives. It turns out that the material of this section is normally used only to obtain a quick general idea of what the Lie derivative of something is or is like, rather than being used directly in proofs. For this reason, the intuitive idea is more important than the details. In order to save time, we shall take advantage of this circumstance as follows: Although the claims of this section will be both precise and true, we shall merely sketch the proofs.

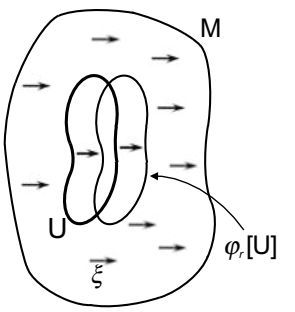


Fix, once and for all, a C^p ($p \geq 2$) manifold M based on Banach space E, together with C^{p-1} $E-$field ξ on M. Fix also positive number a and open subset U of M such that the following property is satisfied: for each point p of U, the maximal integral curve of ξ with initial value p is defined at least on $(-a, a)$. Thus, in the example of the figure, the "U" shown would do the job. However, we could not in this example choose $U = M$, for the maximal integral curve with initial point p shown is not defined for $r-$values ip to a. Denote by I the open interval $(-a, a)$. Next, denote by φ the following mapping from $I \times U$ to M: For (r, p) in $I \times U$ so r is in $(-a, a)$ and p is in U), $\varphi(r, p)$, is the point $\gamma(r)$ of M, where γ is the maximal integral curve of ξ with initial point p. [The condition on U and a above is now seen as having been necessary in order that this φ be well-defined.] Thus, for example, we have $\varphi(0, p) = p$ for every p in U, and $\varphi(r, \varphi(r', p)) = \varphi(r + r', p)$ whenever both sides are defined.

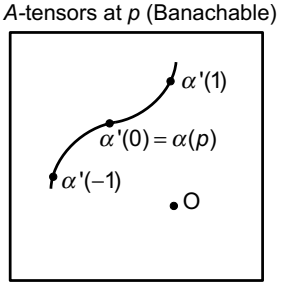
We next note that, since both I and U are manifolds (as open subsets of manifolds), so is $I \times U$. Hence, φ is a mapping from one C^p manifold ($I \times U$) to another (M). We claim: This φ is a C^p mapping of manifolds. To prove this, one first passes to a chart: Let O be an open subset of $E, \hat{\xi}$ a C^{p-1} mapping from O to E, U an open subset of O as above, and $\hat{\varphi}$ the mapping from $I \times U$ to O as above. We first show continuity. We have $|\hat{\varphi}(r', p') - \hat{\varphi}(r, p)| = |\hat{\varphi}(r', p') - \hat{\varphi}(r, p')| + |\hat{\varphi}(r, p') - \hat{\varphi}(r, p)|$. Set $c = \text{lub}\,|\hat{\xi}|$ and $d = \text{lub}\,|\text{D}\hat{\xi}|$. The first term on the right above is $|\hat{\gamma}(r') - \hat{\gamma}(r)|$, where $\hat{\gamma}$ is the maximal integral curve with initial point p'. Hence, this first term is less than or equal to $c\,|r' - r|$. For the second term, set $s = |p' - p|$. Partition the open interval $(0, r)$ into n segments of maximum length ϵ as on page 63. Let $s_1, s_2, \ldots, s_n$ and μ and ν be as on that page. Then we have $s_1 \leq \mu s + \nu$, whence $s_n \leq \mu s_1 + \nu \leq \mu^2 s + \nu(\mu + 1)$, etc., to $s_n \leq \mu^n s + \nu(\mu^{n-1} + \ldots + 1)$. That is to say, $s_n \leq \epsilon^{end} s + \epsilon c/4(e^{\epsilon nd} - 1)$. Letting n go to infinity, keeping the efficiency greater than $1/2$, we obtain $|\hat{\varphi}(r, p') - \hat{\varphi}(r, p)| \leq e^{2rd} s$. Thus, we have obtained $|\hat{\varphi}(r', p') - \hat{\varphi}(r, p)| \leq c\,|r' - r| + e^{2rd}\,|p' - p|$, from which continuity follows immediately. We next show differentiability of $\hat{\varphi}$ at $(0, p)$. We have $|\hat{\varphi}(r', p') - \hat{\varphi}(0, p) - (p' - p) - r'\hat{\xi}(p)| = |\hat{\varphi}(r', p') - p' - r'\hat{\xi}(p)| \leq |\hat{\varphi}(r', p') - p' - r'\hat{\xi}(p')| + r'\,|\hat{\xi}(p') - \hat{\xi}(p)| \leq |\hat{\varphi}(r', p') - p' - r'\hat{\xi}(p')| + r' d\,|p' - p|$, where we used $\hat{\varphi}(0, p) = p$ in the first step, and the mean value theorem in the third. But we have already shown that the first term on the right is tangent at $r' = 0, p' = p$ while the second is obviously tangent. Hence, $\hat{\varphi}$ is differentiable at $(0, p)$, and $D\hat{\varphi}(0, p)$ is the mapping from $\mathbb{R} \times E$ to E which takes (b, x) to $x + b\hat{\xi}(p)$. Along similar lines, one shows that $\hat{\varphi}$ is differentiable everywhere, and then that it is C^p.

The result above, smoothness of $\hat{\varphi}$, is the basic result on the dependence of solutions of differential equations on the initial point and on the parameter. It states that "where you are in M after going amount r along the integral curve with initial point p depends smoothly on both r and p". Indeed, without such a result it is unlikely that differential equations would be of much use in physics. Think of M as representing the "space of states of some physical system", and of the parameter in the integral curves as representing "time". Then the integral curves represent "the evolution of the system through a

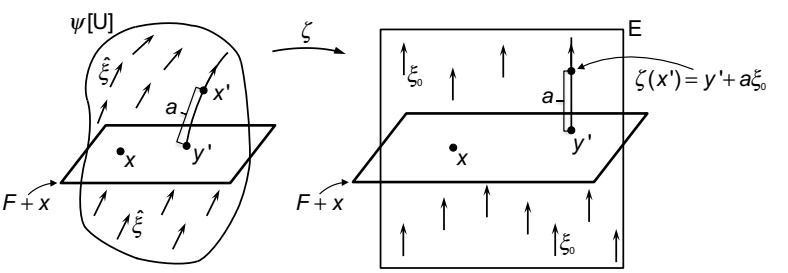
succession of physical states with time". The initial point is "the state of the system at time zero". Now, one cannot avoid some small error in assigning to one's actual physical system an initial point (for, e.g., meters can only be read to a certain accuracy). If one is to do physics sensibly, it had better be true that these small errors in the assignment of initial point result only in small errors in one's prediction of what the system will be like at later times. That is to say, $\hat{\varphi}$ had better be continuous.



Let M, p, E, ξ, U, a, and φ be as above. Let, for each r in $(-a, a)$, φ_r be the mapping from U to M with action $\varphi_r(p) = \varphi(r, p)$. This φ_r, then, is the mapping which "slides each point of U an amount r along its integral curve". Each φ_r is a diffeomorphism from U to $\varphi[U]$ (for $\varphi_r \cdot \varphi_{-r}$ and $\varphi_{-r} \cdot \varphi_r$ are the identity wherever defined). Now let, in addition, α be some C^{p-1} $A-$field on M. Restricting α to the open submanifold U of M, we obtain an $A-$field $\underset{\sim}{\alpha}$ on U. Since φ_r is a diffeomorphism from U to $\varphi_r[U]$, it takes this $\underset{\sim}{\alpha}$ to some $A-$field on $\varphi - r[U]$. Next, fix a point p of U. Then for all sufficiently small r, p will be in $\varphi_r[U]$. In particular, the $A-$field we have defined on $\varphi_r[U]$ will determine an $A-$tensor at p. That is to say, we have obtained for each sufficiently small r, an $A-$tensor at p, which we denote $\alpha'(r)$. We can regard this α', then, as a curve in the Banachable space of $A-$tensors at p. [Intuitively, we "drag our original $A-$field α along by ξ, continually evaluating at p during the dragging, thus obtaining a one-parameter family of $A-$tensors at p".] Note, e.g., that $\alpha'(0) = \alpha(p)$ (since φ_0 is the identity), i.e., our curve has initial point which is just the value of the original $A-$field at p. Since φ is C^p, this curve is also C^p.



We now claim: The tangent vector to this curve α' at 0 is precisely $\mathcal{L}_\xi \alpha$, evaluated at p. That is to say, $\mathcal{L}_\xi \alpha$ at p is just "minus the rate of change of α as it is dragged along by ξ and evaluated at p". To prove this claim, one again passes to a chart. Let us suppose for a moment that one could manage to find a chart, containing p, in which the component $\hat{\xi}$ of ξ is constant, say ξ_0 (in E). Then, with respect to this chart, the component of $\mathcal{L}_\xi \alpha$ at p would be just $D\hat{\alpha}(\xi_0)$ (since other term in the definition of the Lie derivative involves $D\hat{\xi}$, which here vanishes: x is the point of E corresponding to p via our chart). In terms of this chart, however, our curve in the tensor space A is just $\alpha'(r) = \hat{\alpha}(x - r\xi_0)$ (for, in this chart, $\hat{\xi}$, is constant, whence its integral curves are straight lines). Thus, we have in this case that the component of the tangent to the curve α' is precisely the component of $\mathcal{L}_\xi \alpha$ at p, i.e., we have that our claim is true.



There remains, therefore, only to show that there exists a chart containing p with respect to which $\hat{\xi}$ has constant component. We proceed as follows. First, choose any chart, U, ψ, containing p, and let $\hat{\xi}$ be the component of ξ. Next, choose, a subspace F of E complementary to the vector $\xi_0 = \hat{\xi}(x)$. We introduce a mapping ζ from $\psi[U]$ to E as follows: For x' in ψU set $\zeta(x') = y' + a\xi_0$, where y' is the point of $\psi[U]$ at which the

integral curve through x' meets $x + F$, and a is the parameter-difference along this curve from y' to x'. [By choosing U sufficiently small, this mapping will be well-defined.] Thus, for example, $\zeta(x) = x$. This is a C^p mapping with C^p inverse (by smoothness of φ and the inverse function theorem). Now set $\psi' = \zeta \cdot \psi$. Then we obtain a chart on M with respect to which the component of ξ is constant (since, in terms of this new chart, the integral curves of ξ are "straight lines", by construction).

We conclude, therefore, that $\mathcal{L}_\xi \alpha$ can be interpreted as the operation of "taking the rate of change of the field at each point under the 'sliding along' induced by motion along ξ". It is, for example, obvious from this interpretation that the Lie derivative of a natural tensor is zero, since natural tensors are invariant under diffeomorphisms. The other properties of Lie derivatives can be seen in a similar way.

22. Exterior Derivatives

We now consider the second "derivative-like notion" on a manifold. It turns out that the exterior derivative is applicable to only certain types of fields. We begin, therefore, with the study of these fields and their algebra.

Let E be any Banach space. Let n be any non-negative integer, and let σ be the following natural element of the tensor space $\mathcal{L}(\mathcal{L}(E,\dots,E;\mathbb{R});\mathcal{L}(E,\dots,E;\mathbb{R}))$ (with n E's on each side): For κ in $\mathcal{L}(E,\dots,E;\mathbb{R})$, $\sigma(\kappa)$ is that element of $\mathcal{L}(E,\dots,E;\mathbb{R})$ whose action on $x_1,\dots,x_n$ in E is $\sigma(\kappa)(x_1,\dots,x_n) = (1/n!)\Sigma(-1)^s\kappa(x_i,\dots,x_j)$, where the sum extends over all $n!$ permutations of the vectors $x_1,\dots,x_n$, where $x_i,\dots,x_j$ is that permutation, and where s has the value +1 if the permutation is even and -1 if that permutation is odd. [For $n=0,\sigma$ is the identity.]

Example. Set $n=2$. Then $n!=2$, and there are just two permutations of x_1,x_2, namely that which sends this to x_1,x_2 (even), and that to x_2,x_1, (odd). Hence, for κ in $\mathcal{L}(E,E;R)$, $\sigma(\kappa)$ has action $\sigma(\kappa)(x_1,x_2) = (1/2)(\kappa(x_1,x_2)-\kappa(x_2,x_1))$.

Example. Set $n=3$. Then $\sigma(\kappa)(x_1,x_2,x_3) = (1/6)(\kappa(x_1,x_2,x_3)+\kappa(x_2,x_3,x_1)+\kappa(x_3,x_1,x_2)-\kappa(x_2,x_1,x_3)-\kappa(x_3,x_2,x_1)-\kappa(x_1,x_3,x_2))$. This σ will be called the *antisymmetrizer*, and its action that of *taking the antisymmetric part*.

We next claim that this σ has the following property: $\sigma\cdot\sigma=\sigma$, i.e., the antisymmetric part of antisymmetric part is the antisymmetric part. To prove this, let κ be in $\mathcal{L}(E,\dots,E;\mathbb{R})$. Then $\sigma(\sigma(\kappa))(x_1,\dots,x_n) = (1/n!)\Sigma(-1)^s$ $\sigma(\kappa)(x_i,\dots,x_j) = (1/n!)\Sigma(-1)^s[(1/n!)\Sigma(-1)^{s'}\kappa(x_{i'},\dots,x_{j'})]$, in an obvious notation, where we have used the definition twice. The double sum on the right consists of $(n!)^2$ terms, each of which is $(1/n!)^2$ times plus or minus κ applied to some permutation of $x_1,\dots,x_n$, plus if that permutation is even, minus if odd. But there are only $n!$ permutations of $x_1,\dots,x_n$. Thus, each permutation occurs $(n!)$ times in this sum. Combining the terms corresponding to the same permutation, we obtain $(1/n!)^2\ \Sigma(-1)^s\kappa(x_i,\dots,x_j)(n!)$, where the sum is over permutations, and where the last $n!$ arises because $n!$ terms of the original sum yield a single term of this one. But this last is precisely $\sigma(\kappa)(x_1,\dots,x_n$. Thus, we have $\sigma(\sigma(\kappa))(x_1,\dots,x_n) = \sigma(\kappa)(x_1,\dots,x_n)$, whence since the x's and κ are arbitrary, we have $\sigma\cdot\sigma=\sigma$. [This argument is much easier than it looks. One should try it explicitly, e.g., for $n=2$.]

An element of $\mathcal{L}(E,\dots,E;\mathbb{R})$, ($nE$'s), κ is called an *n–form* if $\sigma(\kappa)=\kappa$. [A *0–form* is a number.] For example, every element of $\mathcal{L}(E;\mathbb{R})$ is a 1–form. From the result of the previous paragraph we have: for any κ in $\mathcal{L}(E,\dots,E;\mathbb{R})$, $\sigma(\kappa)$ is an n–form (for $\sigma(\sigma(\kappa)) = \sigma(\kappa)$). We also have the following property of n–forms: For κ an n–form, $x-1,\dots,x_n$ in E, and $x_i,\dots,x_j$ a permutation of these, $\kappa(x_1,\dots,x_n) = (-1)^s\kappa(x_i,\dots,x_j)$, where s is +1 if $x_i,\dots,x_j$ is an even permutation of $x_1,\dots,x_n$, and -1 if odd. [Proof: This is immediate from the definition of σ, for precisely the same terms appear in $\sigma(\kappa)(x_1,\dots,x_n)$ as in $\sigma(\kappa)(x_i,\dots,x_j)$, although all the terms of the latter will differ in sign from those of the former if $s=-1$.]

Our concern, for the moment, is the algebra of these n–form (a subject usually called exterior algebra). We next introduce a certain product between these forms. Fix non-negative integers n and $n!$ and let τ be the following natural element of $\mathcal{L}(\mathcal{L}(E,\dots,E;\mathbb{R}),\mathcal{L}(E,\dots,E;\mathbb{R});\mathcal{L}(E,.$ (with n,n' and $n+n'$ E's, respectively): For κ in $\mathcal{L}(E,\dots,E;\mathbb{R})$ (n E's) and κ' in $\mathcal{L}(E,\dots,E;\mathbb{R})$ (n' E's), let $\tau(\kappa,\kappa)$ be that element of $\mathcal{L}(E,\dots,E;\mathbb{R})$ ($n+n'$ E's) with $\tau(\kappa,\kappa')(x_1,\dots,x_{n+n'}) = \kappa(x_1,\dots,x_n)\kappa'(x_{n+1},\dots,x_{n+n'})$. Thus, τ simply "takes the product of the multilinear mappings" in the natural way. Of course, if κ and κ' happen to be n– and n'–forms, respectively, it will not in general be true that $\tau(\kappa,\kappa')$ is an $n+n'$)–form. [For example, let κ and κ' be nonzero 1–forms. Then $\tau(\kappa,\kappa')(x_1,x_2) = \kappa(x_1)\kappa'(x_2)$, which

will not be equal to $-\tau(\kappa,\kappa')(x_2,x_1)$, whence $\tau(\kappa,\kappa')$ will not be a 2−form.] We can, however, obtain an $(n+n')$−form by applying σ. Thus: for κ and κ' $n-$ and $n'-$forms, respectively, the $(n+n')$−form $\sigma(\tau(\kappa,\kappa'))$ is written $\kappa\wedge\kappa'$ and is called the *wedge product* of κ and κ'. [For either n or n' zero, one multiples the form by the number.]

We have now defined forms, together with a product operation on them. We wish next to obtain properties, of which there are three. First, we have that the wedge product is linear in each factor, i.e., $(\kappa+\underline{\kappa})\wedge\kappa' = \kappa\wedge\kappa'+\underline{\kappa}\wedge\kappa'$ and $\kappa\wedge(\kappa'+\underline{\kappa'}) = \kappa\wedge\kappa'+\kappa\wedge\underline{\kappa'}$, where κ and $\underline{\kappa}$ are n−forms and κ' and $\underline{\kappa'}$ are n'−forms. These facts, together with the fact that the sum of two n−forms is an n−form, are all immediate, since everything in sight linear. Second, we have; For κ,κ', and κ'', $n-$, $n'-$, and $n''-$forms, respectively, $(\kappa\wedge\kappa')\wedge\kappa'' = \kappa\wedge(\kappa'\wedge\kappa'')$ (each side an $(n+n'+n'')$−form). [Proof: First note that, for any κ and κ', $\sigma\cdot\tau(\sigma(\kappa),\kappa') = \sigma\cdot\tau(\kappa,\kappa')$, for, applying both sides to $(x_1,\ldots,x_{n+n})$ the innermost "σ" on the left will antisymmetrize over the first n "x's" in each term, which is unnecessary, since the outermost "σ" already antisymmetrizes over all "x's". Next, note that $\tau(\tau(\kappa,\kappa'),\kappa'') = \tau(\kappa,\tau(\kappa',\kappa''))$, which is immediate from the definition (applying to $x_1,\ldots,x_{n+n'+n''}$, κ will get the first n x's, κ' the next n', and κ'' the last n'', using either side of the equation). We now have $(\kappa\wedge\kappa')\wedge\kappa'' = \sigma\cdot\tau(\sigma\cdot\tau(\kappa,\kappa'),\kappa'')) = \sigma\cdot\tau(\tau(\kappa,\kappa',\kappa'') = \sigma\cdot\tau(\kappa,\tau(\kappa',\kappa'')) = \sigma\cdot\tau(\kappa,\sigma\cdot\tau(\kappa',\kappa'')) = \kappa\wedge(\kappa'\wedge\kappa'')$, where we used definitions in the first and fifth steps, our first observation above in the second and fourth, and our our second observation in the third.] Thus, the wedge product is associative. Finally, we claim a sort of commutativity: For κ and κ' $n-$ and $n'-$forms, respectively, $\kappa\wedge\kappa' = (-1)^{nn'}\kappa'\wedge\kappa$. [Proof: We have $\tau(\kappa,\kappa')(x_1,\ldots,x_{n+n'}) = \kappa(x_1,\ldots,x_n)\kappa'(x_{n+1},\ldots,x_{n+n'})$, and $\tau(\kappa',\kappa)(x_1,\ldots,x_{n+n'}) = \kappa'(x_1,\ldots,x'_n)\kappa(x_{n'+1},\ldots,x_{n'+n}) = \kappa(x_{n'+1},\ldots,n_{n'+n'})\kappa'(x_1,\ldots,x'_n)$. The order of the x's in the expression on the right in the second formula is an odd permutation of that in the first formula if and only if both n and n' are odd. Hence, $\sigma(\tau(\kappa,\kappa')) = (-1)^{nn'}\sigma(\tau(\kappa',\kappa))$.]

This, then, is the set-up: We have n−forms, and on them a bilinear, associative, and "more or less commutative" product. This algebra in tensor spaces goes over immediately to fields on a manifold M. Thus, an *n−form* (field) on M is an $\mathcal{L}(E,\ldots,E;\mathbb{R})$−field whose component with respect to any chart at any point is an n−form. Since the wedge product is a natural operation, one can take the wedge product of n−forms on M. Since our three properties hold pointwise, they hold for the fields.

Let M be a C^p $(p\geq 2)$ manifold based on Banach space E. The exterior derivative, which we now define, will be applicable only to (C^{p-1}) forms on M. Let, then, κ be a C^{p-1} n−form on M. Passing to a chart, we have an open subset O of E, and a C^{p-1} mapping $\hat{\kappa}$ from O to the tensor space $\mathcal{L}(E,\ldots,E;\mathbb{R})$, where, for each $x, \sigma(\hat{\kappa}(x)) = \hat{\kappa}(x)$ (i.e., where each $\hat{\kappa}(x)$ is an n−form in our tensor space). Now fix x in O, and consider $D\hat{\kappa}(x)$, an element of $\mathcal{L}(E;\mathcal{L}(E,\ldots,E;\mathbb{R}))$.

We next note that there is a natural isomorphism from $\mathcal{L}(E;\mathcal{L}(E,\ldots,E;\mathbb{R}))$ (n E's on the right) to $\mathcal{L}(E,\ldots,E;\mathbb{R})$ ($n+1$ E's), namely, that which sends τ in the former to that element of the latter which acting on $x_1,\ldots,x_{n+1}$ in E, produces the real number $\tau(x_1)(x_2,\ldots,x_{n+1})$. In what follows, action of this isomorphism will be assumed implicitly when appropriate. Thus, in particular, we may regard $D\hat{\kappa}(x)$ as an element of $\mathcal{L}(E,\ldots,E;\mathbb{R})$ ($n+1$ E's). Hence, applying the antisymmetrizer, $\sigma(D\hat{\kappa}(x)$ is an $n+1$)−form. We now claim that, in fact, this construction yields a tensor at the point p of M. That is, let U,ψ and U',ψ' be two charts, with p in $U\cap U'$. Then, denoting by $\hat{\kappa}$ and $\hat{\kappa'}$ the respective components of κ, we have that, for $x_1,\ldots,x_n$ in E, $\hat{\kappa'}(x')(\iota(x_1\ldots,\iota(x_n)) = \hat{\kappa}(x)(x_1,\ldots,x_n)$, where $x=\psi(p)$, $x'=\psi'(p)$, and where $\iota = D(\psi'\cdot\psi^{-1})(x)$. Taking the derivative of each side of this equation (with respect to x), and applying to an arbitrary vector y in E, we have $D\hat{\kappa'}(x')(\iota(y))(\iota(x_1),\ldots,\iota(x_n)) + \hat{\kappa'}(x')(D\iota(y)(x_1),\ldots,\iota(x_n)) + \ldots + \hat{\kappa'}(x')(\iota(x_1),\ldots,D\iota(y)(x_n)) = D\hat{\kappa}(x)(y)(x_1,\ldots,x_n)$. Now apply, to each side of this equation, the antisymmetrizer (over the $(n+1)$ vectors $y,x_1,\ldots,x_n$). Then all the terms except the first on the left vanish (for, e.g., the second term, we have $D\iota(y)(x_1) = D\iota(x_1)(y)$, by symmetry of mixed partials, while, in the sum resulting from application of the antisym-

metrier, for each containing $D\iota(y)(x_1)$ there will be another term identical except for sign but with this replaced by $D\iota(x_1)(y)$. Hence, the terms will all cancel in pairs.) Thus, we obtain $\sigma(D\hat{\kappa}'(x'))(\iota(y), \iota(x_1), \ldots, \iota(x_n)) = \sigma(D\hat{\kappa}(x))(y, x_1, \ldots, x_n)$. But this is precisely the statement that $(\sigma(D\hat{\kappa}'(x')); U', \psi') \approx (\sigma(D\hat{\kappa}(x)); U, \psi)$. Thus, we obtain an $(n+1)-$form at p.

Repeating the above at each point of M, we conclude: Given a C^{p-1} $n-$form (failed) on M, we obtain as above a C^{p-2} $(n+1)-$form (field on M). This latter is called the *exterior derivative* of κ, written $d\kappa$.

Example. Let $n = 0$. The $0-$form κ is just a scalar field. The instructions above read in this case: The exterior derivative d is that 1−form (i.e., $\mathcal{L}(E;\mathbb{R})-$field such that, passing to a chart and letting y in E be arbitrary, $\widehat{d\kappa}(x)(y) = D\hat{\kappa}(x)(y)$ (antisymmetrization being unnecessary for 1−forms). But the right side of the above is precisely the directional derivative of κ in the $y-$direction. That is to say, $d\kappa$ is what one in elementary calculus calls the gradient. We conclude: On $0-$forms, the exterior derivative is the gradient.

Example. Let $n = 1$. Then, for κ a 1−form on M, the exterior derivative has the following expression in terms of a chart: $(d\hat{\kappa}(x))(y, x_1) = D\hat{\kappa}(x)(y)(x_1) - D\hat{\kappa}(x)(x_1)(y))$. That is to say, one takes "half the derivative of the x_1-component of $\hat{\kappa}$ in the $y-$direction minus half the derivative of the $y-$component of $\hat{\kappa}$ in the x_1-direction". These instructions will be recognized as those which, in elementary calculus, yield the curl. On the 1−forms, therefore, the exterior derivative is the curl.

On higher forms, the exterior derivative is a sort of "generalized curl".

As usual, we now want to determine the various properties of the exterior derivatives. There are four. First, we have: For κ and λ $n-$forms, $d(\kappa+\lambda) = d\kappa + d\lambda$. Second: For κ and κ' $n-$ and $n'-$forms, respectively, $d(\kappa\wedge\kappa') = d\kappa\wedge\kappa' + (-1)^n \kappa\wedge d\kappa'$. This is a sort of "Leibniz rule". Note that this equation makes sense, each side being an $(n+n'+1)-$form. [Proof: In terms of a chart, the left side, applied to $y, x_1, \ldots, x_{n+n'}$, is $\sigma(D(\hat{\kappa}(x_1, \ldots, x_n)\,\hat{\kappa}'(x_{n+1}, \ldots, x_{n+n'}))(y)) = \sigma(D\hat{\kappa}(y)(x_1, \ldots, x_n)\,\hat{\kappa}'(x_{n+1}, \ldots, x_{n+n'})) + \sigma(\hat{\kappa}(x_1 \ldots, x_n)\, D\hat{\kappa}'(y)(x_n, \ldots, x_{n+n'}))$. But the two terms on the right in this formula are the chart-representations of the two terms on the right in our claimed equation.]

Example. Let $n = 0$, $n' = 1$. Then κ is a scalar field, κ' a 1−form, and our property above becomes $d(\kappa\kappa') = d\kappa \wedge \kappa' + \kappa d\kappa'$. This will be recognized as a formula from elementary vector calculus: The curl of a function times a vector is given by the gradient of the function cross the vector plus the function times the curl of the vector.

The third property is this: For any $n-$form κ, $dd\kappa = 0$). [Proof: In terms of a chart, the left side, applied to $z, y, x_1, \ldots, x_n$, is $\sigma D(D\hat{\kappa}(y))(z)(x_1, \ldots, x_n) = \sigma(DD\hat{\kappa}(y)(z)(x_1, \ldots, x_n))$, where the antisymmetrizer is over $z, y, x_1, \ldots, x_n$. Since mixed partial are symmetric, the antisymmetrizer annihilates.]

Example. Set $n = 0$ in the last property. Then, in elementary calculus terminology, we have: The curl of the gradient of a scalar field vanishes. This, of course, is true.

The final property relates Lie and exterior derivatives. For this property, we need a bit of notation. For xi an $E-$tensor and κ an $n-$form, we write $\xi \cdot \kappa$ for the $(n-1)-$form whose action on $x_1, \ldots, x_{n-1}$ is $(\xi \cdot \kappa)(x_1, \ldots, x_{n_1}) = \kappa(\xi, x_1, \ldots, x_{n_1})$. Similarly for fields. [This, of course, is the action of a natural operation.] Thus, for example, we have $\xi \cdot (\xi \cdot \kappa) = 0$. The final property is this: For ξ an $E-$field and κ and $n-$form (field), $(n+1)\xi \cdot d\kappa - \mathcal{L}_\xi \kappa + n\, d(\xi \cdot \kappa) = 0$. [Note that this is well-defined, each term on the left being an $n-$form.] [Proof: Passing to a chart and applying to $x_1, \ldots, x_n$, $\xi \cdot d\kappa$ becomes $\sigma(D\hat{\kappa}(\hat{\xi})(x_1, \ldots, x_n))$ where the antisymmetrizer is over $\hat{xi}, x_1, \ldots, x_n$; $\mathcal{L}_\xi \kappa$ becomes $D\hat{\kappa}(\hat{\xi}(x_1, \ldots, x_n) + \hat{\kappa}(D\hat{\xi}(x_1, \ldots, x_n)) + \ldots + \hat{\kappa}(x_1, \ldots, D\hat{\xi}(x_n))$; and $d(\xi \cdot \kappa)$ becomes $\sigma(D(\hat{\kappa}(\hat{\xi}, x_2, \ldots, x_n))(x_1))$, where the antisymmetrizer is over $x_1, \ldots, x_n$. Expanding the antisymmetrizer in all three expressions, and the last expression using the Leibnitz rule, one sees that all the terms in the claimed combination cancel.]

Example. Setting $n = 0$, our last property becomes $\xi \cdot d\kappa - \mathcal{L}_\xi \kappa = 0$ for κ a $0-$form. But this formula is true, for each of the two terms on the left has been seen to be the directional derivative of κ in the $\xi-$direction.

In the finite-dimensional case, the last property above provides a convenient way of defining the exterior derivative, without reference to charts. Rewrite the the formula of that property in the form $\xi \cdot d\kappa = 1/(n+1)\,\mathcal{L}_\xi \kappa - n/(n+1)\,d(\xi \cdot \kappa)$. Then, for $n = 0$, the right side is known (once one has Lie derivatives), and so this formula can be used to define the exterior derivative of 0−form. Once exterior derivatives have been defined on forms up to $(n-1)$, the right side of this formula, for κ an n−form, is known, whence this formula can be used to define the exterior derivative of an n−form. By induction, then, one obtains the definition of exterior derivative on any form. A similar program could be carried out in the infinite-dimensional case, but, as usual, additional complications arise from the feature that manifolds may not admit non-trivial fields globally.

23. Derivative Operators

We come, finally, to the third "type of derivative".

Fix a C^p ($p \geq 2$) manifold M based on Banach space E. Fix a point p of M. Consider now pairs, $(\Gamma; U, \psi)$, where U, ψ is a chart containing p, and Γ is an element of $\mathcal{L}(E, E; E)$ which is symmetric (i.e., which satisfies, for any y, y' in E, $\Gamma(y, y') = \Gamma(y', y)$). Given two such, we write $(\Gamma; U, \psi) \approx (\Gamma'; U', \psi')$ provided the following equation holds: For any vectors y and z in E, $\Gamma'(\iota(y), \iota(z)) = \iota(\Gamma(y, z)) + D\iota(x)(y, z)$, where, as usual, $\iota = D(\psi' \cdot \psi^{-1})$, and $x = \psi(p)$. Each side of this equation is an element of E. it should be noted that this equation is different from the corresponding equation for an $\mathcal{L}(E, E; E)$ tensor at p: Indeed, were the second term on the right above omitted, we would have precisely the tensor relation. The things we are now defining are "geometrical objects" which are somewhat like, but not precisely the same as, tensors.

We claim that "$\approx$" is an equivalence relation. Clearly, we have $(\Gamma; U, \psi) \approx (\Gamma; U, \psi)$. Suppose next that $(\Gamma; U, \psi) \approx (\Gamma'; U', \psi')$. Then we have $\Gamma'(\iota(y), \iota(z)) = \iota(\Gamma(y, z)) + D\iota(x)(y, z)$, from which it follows, we claim that $\Gamma(\iota^{-1}(y), \iota^{-1}(z)) = \iota^{-1}(\Gamma(y, z) + D(\iota^{-1})(x')(y, z)$, [Proof: Expand $D(\iota^{-1}$ in the second formula, and replace y by $\iota(y)$ and z by $\iota(z)$.] But this second formula is precisely the statement that $(\Gamma'; U', \psi') \approx (\Gamma; U, \psi)$. Finally, that $(\Gamma; U, \psi) \approx (\Gamma'; U', \psi')$ and $(\Gamma'; U', \psi') \approx (\Gamma''; U'', \psi'')$ implies $(\Gamma; U, \psi) \approx (\Gamma''; U'', \psi'')$ is, similarly, an easy exercise in algebra.

An equivalence class of such pairs is called a *connection* at p. A connection field on M is a mapping which associates with each point p of M a connected at p. The component of a connection field with respect to a chart is the corresponding mapping from $O = \psi[U]$, an open subset of E, to $\mathcal{L}(E, E; E)$. A connection field is said to be C^{p-2} if its component is C^{p-2} for every admissible chart. [Note that we here go from p all the way down to $(p-2)$. The reason is that "$D\iota$", the second derivative of the chart-mappings, appears in the formula for component-change under chart-change.] A C^{p-2} connection field on M is normally called a *derivative operator* on M.

The interest in the notion of a derivative operator stems from the feature that such an operator permits one to take the derivative of an arbitrary tensor field on M, again obtaining a tensor field. We next see how this comes about. Let α be a C^{p-1} $A-$field on M, and fix a derivative operator on M. Choose a chart, and let $\hat{\alpha}$ and $\hat{\Gamma}$ be the components of α and the derivative operator, respectively. Now let x be any point of $O = \psi[U]$, and y any vector in E, and consider the free $A-$tensor $D\hat{\alpha}(x)(y) - \kappa(\widehat{\Gamma}(x)(y), \hat{\alpha}(x))$, where κ is the element of $\mathcal{L}(\mathcal{L}(E; E), A; A)$ defined at the top of page 59. For each y in E, the above is a free $A-$tensor, clearly linear in y. Hence, the above can be regarded as a mapping from E to A, i.e., as a free $\mathcal{L}(E; A)-$tensor. Now change the chart. We have the formulae for how $\hat{\kappa}$ and $\hat{\Gamma}$ change, and so can write down how the free $\mathcal{L}(E; A)-$tensor above changes. We claim: It behaves like a $\mathcal{L}(E; A)-$tensor at p. The calculation is identical with that for Lie derivatives: One picks up a "$D\iota$" term from "$D\hat{\alpha}$", and also a "$D\iota$" term from $\widehat{\Gamma'}$ (by the formula for the component-change under chart-change for a connection). The expression above has been adjusted so that these terms cancel. Thus, we obtain an $\mathcal{L}(E; A)-$tensor at p, which we write $\nabla\alpha(p)$. Repeating for each point of M, we obtain a C^{p-2} $\mathcal{L}(E; A)-$ field on M, $\nabla\alpha$. This field is called the *derivative* of α (with respect to our given derivative operator).

Example. Let α be a scalar field. Then $\nabla\alpha$ must be a $\mathcal{L}(E; \mathbb{R})-$field, i.e., a $1-$form. In this case, the "$\kappa-$term" above vanishes, and we have $(\widehat{\nabla\alpha})(y) = D\hat{\alpha}(y)$. But this is precisely the formula for the exterior derivative. Hence, the result of application of a derivative operator to a scalar field is to take its exterior derivative.

Thus, the derivative, $\nabla\alpha$, of a C^{p-1} $A-$field α is a C^{p-2} $\mathcal{L}(E; A)-$field. This is of course what one might have expected: The "E" refers to the possible directions in M

along which the derivative might be taken.

We obtain a few properties of derivatives via derivative operators. Clearly, $\nabla(\alpha+\alpha') = \nabla\alpha + \nabla\alpha'$. Next, let τ be a natural tensor field. Then, for its component with respect to a chart, $\hat{\tau}$ is a constant mapping. Hence, $D\hat{\tau} = 0$. Furthermore, as we saw in Sect. 19, $\kappa(\hat{\Gamma}, \hat{\tau}) = 0$ in this case. from the formula above for the derivative, therefore, we conclude that $\nabla\tau = 0$. The derivative of a natural field vanishes. We next note that, for ρ any $\mathcal{L}(A_1, \ldots, A_n; B)-$field, and $\alpha_1, \ldots, \alpha_n$ $A_1, \ldots, A_n-$fields, we have $\nabla(\rho(\alpha_1, \ldots, \alpha_n)) = (\nabla\rho)(\alpha_1, \ldots, \alpha_n) + \rho(\nabla\alpha_1, \ldots, \alpha_n) + \ldots + \rho(\alpha_1, \ldots, \nabla\alpha_n)$. it follows immediately that the derivative satisfies the Leibnitz rule for any natural operation on tensor fields. Finally, we show that "mixed derivatives commute", at least when applied to scalar fields. Let f be a scalar field. Then $\nabla\nabla f$ is a $\mathcal{L}(E, E; R)-$field. We claim that this $\nabla\nabla f$ is symmetric (i.e., $\nabla\nabla f(x_1, x_2) = \nabla\nabla f(x_2, x_1)$). Indeed, choosing a chart, we have $\widehat{\nabla f}(x_1) = D\hat{f}(x_1)$, whence $\widehat{\nabla\nabla f}(x_1, x_2) = DD\hat{f}(x_1, x_2) - Df(\widehat{\Gamma}(x_1, x_2))$. But the first term is symmetric in x_1 and x_2 since mixed partials commute, while the second is symmetric since $\widehat{\Gamma}$ is.

Again, in the finite-dimensional case one normally defines derivatives operators by their properties, rather than via charts, a method which seems to be more awkward in the infinite-dimensional case.

Derivative operators are a rather "brute force" way of making available a derivative. One simply introduces what one needs (a connection field) to be able to eliminate "$D\iota-$terms" when taking component-derivatives. As one might expect, then, derivative operators will normally not be very useful unless one arises naturally from what one has available to him already. It turns out that there is one situation in which such an operator does appear naturally: in the presence of a metric. Thus, it is normally when one deals with manifolds-with-metrics, i.e., with what is called Riemannian geometry, that derivative operators play the important role. In a similar way, Lie derivatives are not normally very useful unless one happens to have some natural vector field around.

Finally, we remark that there are numerous relationships between Lie derivatives, exterior derivatives, and derivatives via derivative operators. In particular, the former two can be expressed in terms of the latter.

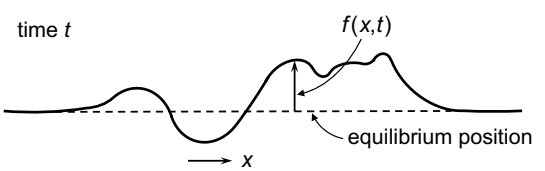
24. Derivatives: Summary

The table below summarizes selected features of the three types of derivative we have discussed.

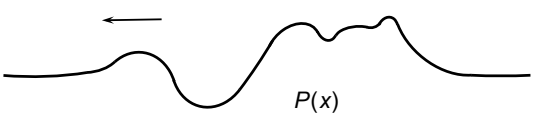
Name	Lie Derivative	Exterior Derivative	Derivative
Applies to	Any field α	Forms only	Any field α
Requires	$E-$field ξ	Nothing	Derivative operator ∇
Symbol	$\mathcal{L}_\xi \alpha$	$\mathrm{d}\alpha$	$\nabla\alpha$
Properties	Additive Leibnitz	Additive "Leibnitz"	Additive Leibnitz
On Natural Fields	Gives zero	—	Gives zero

25. Partial Differential Equations

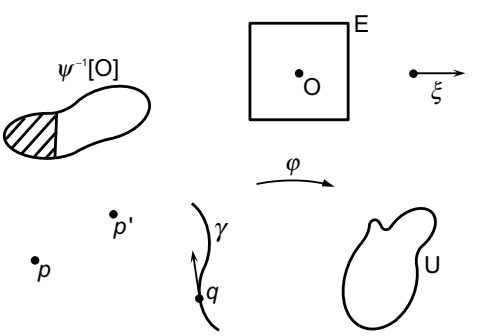
A vector field on a manifold has unique maximal integral curves. As we have seen in Sect. 20, this fact leads, in the finite-dimensional case, to a simple, geometrical statement of the existence and uniqueness of solutions of systems of ordinary differential equations. But the fact is true also in the infinite-dimensional case. It is natural to ask, therefore, whether or not one can, in a similar way, obtain a "geometrization" of certain partial differential equations. It is our goal in this section to look briefly, by means of an example, at this question. Our conclusion, essentially, will be that naive ideas do not work.



We shall be concerned with mappings from $\mathbb{R}^2$ to $\mathbb{R}$, i.e., with real-valued functions of two real variables. We denote such a function f, and the variables t and x. The differential equation whose solutions we wish to study is this: $(\partial_t)^2 f = (\partial_x)^2 f$, where "$\partial_t$" and "$\partial_x$" denote the partial derivatives with respect to t and x, respectively. [This is the equation, e.g., for the propagation of waves on a tightly stretched string. Then "x" denotes position along the string, "t" denotes the time, and "$f(x,t)$" denotes the displacement of the string from its equilibrium position at location x and time t.] This is perhaps the simplest partial differential equation one could invent to test our program: It is linear, hyperbolic, of second order, and in two variables. One of the nice features of this particular equation is that it is easy to write down its general solution. Indeed, let p and q each be functions of one real variable. Then, clearly, the function f with action $f(x,t) = p(x+t)$ is a solution of our equation, as is the function f with action $f(x,t) = q(x-t)$. By linearity, their sum, $p(x+t) + q(x-t)$, is also a solution. [Physically, a solution of the form $p(x+t)$ corresponds to a wave pulse on the string, of shape described by p, which moves to the left along the string without changing its shape. Similarly, a solution of the form $q(x-t)$ corresponds to a pulse moving to the right.] It turns out that, in fact, the most general solution of our partial differential equation is of the form $p(x+t) + q(x-t)$. [A statement which is true, whose proof we omit, but which, as we shall see shortly, could very easily be proven after completion of the program we are here discussing.]



The first step in the geometrization of an ordinary differential equation was to cast it into the language of manifolds. We now wish to do the same thing for the present equation. To this end, let each of E_1 and E_2 be a "space" (eventually, it is hoped, a Banach space) of functions of one real variable, and let E denote their product, $E = E_1 \times E_2$. Consider now a solution f of our partial differential equation. For each real number t_0, let g_{t_0} be the function with action $gt_0(x,t) = f(x,t_0)$ and h_{t_0} the function with action $h_{t_0}(x) = (\partial_t f)(x,t_0)$. [Physically, g_{t_0} describes the "shape" of the string at time t_0, and h_{t_0} the "velocity of motion up or down of each segment of the string at time t_0".] The pair (g_{t_0}, h_{t_0}), then, as a pair of functions of one real variable, should define a point of $E_1 \times E_2$, i.e., a point of E. Since this is the case for each real number t_0, we obtain a mapping γ from the reals to E, with action $\gamma(t_0) = (g_{t_0}, h_{t_0})$. That is, we obtain a curve in E. Of course,



given this curve in E, the original solution f of the partial differential equation can be recovered immediately, for, e.g., $f(x,t)$ is then the first entry of $\gamma(t)$, evaluated at x. In this way, then, we can describe solutions of our partial differential equation by certain curves. [Physically, E is the space of "possible physical states of the string", and the curve describes "the evolution of the string with time through a succession of physical states".]

The general idea is this. We wish to introduce a certain $E-$field on the (eventually, a manifold) E, so constructed that its integral curves are precisely the curves, obtained above, corresponding to solutions of our partial differential equation. There is a simple check one can do to see if this is even close to being possible: One can determine whether a single (e.g., initial) point on such a curve determines the curve (i.e., the solution of our equation) uniquely. That is, we ask: Is it true that, given any (g,h) in E, there is one and only one solution of our partial differential equation, f with $g(x) = f(x,0)$ and $h(x) = (\partial_t f)(x,0)$. In fact, the answer is yes, and the unique solution is given by $f(x,t) = 1/2(g(x+t) + g(x-t)) + 1/2(h(x+t) - h(x-t))$. [The right side is indeed a solution, as we have already observed. Furthermore, as one easily checks directly, the f defined by this equation indeed satisfies $g(x) = f(x,0)$ and $h(x) = (\partial_t f)(x,0)$. That this is the unique f satisfying these conditions is also clear, since every solution f is of the form $p(x+t) + q(x-t)$.] Thus, our "solution curves" do not cross.

We now have our "space" E, and our curves, and we next wish to find a vector field whose integral curves are these curves. But this is quite easy: We simply take the tangent vectors to our curves. Thus, let f be a solution of our partial differential equation, γ the corresponding curve in E, and let $\gamma(0) = (g,h)$. Then γ has action $\gamma(t) = (g_t, h_t)$, with $g_t(x) = f(x,t)$ and $h_t(x) = (\partial_t f)(x,t)$, whence the tangent vector to γ at zero is $(\hat{g}, \hat{h})$, where $\hat{g}(x) = (\partial_t f)(x,0)$, and $\hat{h}(x) = (\partial_t)^2 f(x,0)$. In other words, $g = h$ and $h = g''$ (where prime denotes derivative with respect to the single variable, x on which g depends) this last equation comes from the partial differential equation on f). This little calculation of the tangent vector also tells us what the vector field should be. Let ξ be the mapping from E to E with action $\xi(g,h) = (h,g'')$. [Note what happened here: The tangent vector depended only on (g,h), so it could be represented by some vector field.] Then the integral curves of this γ will be precisely the solution curves of our partial differential equation.

The above discussion is essentially only motivation, for we have omitted the "detail" of specifying precisely which functions are permitted to be in the space E. If, however, this single point could be cleared up, we would have a geometrization of this particular (and then, presumably, also of many similar) partial differential equations. We now concern ourselves with this issue of specifying E.

A natural choice might be this. Choose integer $p \geq 0$, and let E_1 consist of C^p functions g with $\mathrm{lub}\,|g| + \mathrm{lub}\,|g'| + \ldots + \mathrm{lub}\,|g^{(p)}|$ finite, where the norm is this sum. As we have seen, this E_1 is indeed a Banach space. Repeating for E_2, we obtain a Banach space E.

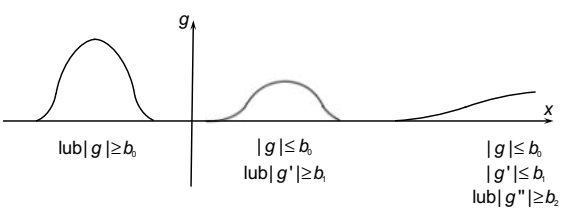
We first observe that, since ξ is to be a mapping from E to E, since ξ is to have action $\xi(g,h) = (h,g'')$, it must be true that, for any (g,h) in E, so is (h,g''). In particular, we must choose $p \geq 2$, in order that "g''" make sense. But even this is not enough, for, having so chosen p, each of h and g'' must be C^p – in general, the latter will be only C^{p-2}. One might hope to get around this problem by, e.g., letting E_1 consist of C^p functions and E_2 of C^{p-2}. Then g will be C^p, whence g'' be C^{p-2}, but this g'' is the second entry of (h,g''), and functions there need only be C^{p-2} (since they are in E_2). This, however, will not work either, for now the "h" in (g,h) will only have to be C^{p-2}, whence "h" is not a suitable first entry in (h,g''). We conclude, therefore, that there seems to be reasonable choices for E_1 and E_2 as spaces of functions of finite differentiability class for which our "ξ" will indeed be a well-defined mapping from E to E. The problem, of course, is that ξ "takes derivatives", which tends to "push us out of the space in which we began".

The obvious solution is to pass to C^∞ functions (for which "take a derivative" does

not get one out of the space). Suppose, then, that we let E_1 consists of C^∞ functions whose value, the value of whose derivative, etc.s all bounded. We could not choose on this E_1 a norm of the form $\text{lub}\,|g| + \text{lub}\,|g'| + \ldots + \text{lub}\,|g^{(p)}|$, for some fixed positive p, for then E_1 would not be complete. [As a general rule, completeness fails when restrictions are imposed on the class of admitted functions which are not suitably incorporated into the norm.] We could, however, set $d(g) = \text{lub}\,|g|/(1 + \text{lub}\,|g|) + 1/2\text{lub}\,|g'|/(1 + \text{lub}\,|g'|) + 1/4\text{lub}\,|g''|/(1 + \text{lub}\,|g''|) + \ldots$, a measure of the "size" of g. [The idea of this expression is that each term, without its numerical factor, is less than one, whence the n^{th} term is less than $1/2^n$, whence the sum always converges.] This is indeed a metric on our set of C^∞ functions, i.e., letting the distance between g and $\tilde{g}$ be $d(g - \tilde{g})$. Furthermore, as one might expect from the appearance of the formula above, our E_1 is complete under this metric. Our E_1 how includes the derivative of each function in it, and is complete. Unfortunately, we have also now introduced a new problem: The "d" above is not a norm, for it fails to be true that $d(ag) = a\,d(g)$ for a real number. [This E_1 is what is called a Frechet space, about which we shall say slightly more later.] The problem, of course, is all those "$1 + \text{lub}''$"s in the denominators.

To summarize, we have found so far that C^p functions will not work for E, because this set of functions does not include the derivative of each of its elements. Furthermore, C functions with "C^p-norm" will not work either (although derivatives are not included) because completeness fails. "Completeness" can be restored for C^∞ functions, but the result is just a metric rather than a norm.

Let us try, then, to see if we can find some norm on C^∞ functions. A natural candidate would be $|g| = a_0\text{lub}\,|g| + a_1\text{lub}\,|g'| + \ldots$ where $a_0, a_1, \ldots$ is some sequence of positive numbers (zero cannot be allowed, for then completeness would fail). We must now concern ourselves with convergence of this infinite sum (a concern that did not arise for "d"). The idea would be to choose the a_n to go to zero sufficiently quickly with n that this sum will converge for every C^∞ g with each term finite. Unfortunately, none exists.



Example. Let $b_0, b_1, \ldots$ be any sequence of positive numbers. Then there exists a C^∞ function g such that each of $\text{lub}\,|g|, \text{lub}\,|g'|$, etc. is finite, but such that $\text{lub}\,|g| \geq b_0$, $\text{lub}\,|g'| \geq b_1$, etc. Indeed, the graph of such a g is zero except for a sequence of "bumps", where, in the first bump, the value of $|g|$ exceeds b_0; in the second, the value does not exceed b_0 but the value of the derivative exceeds b_1; in the third, the value does not exceed b_0 and the value of the derivative does not exceed b_1, but the value of the second derivative exceeds b_2, etc.

Thus, at least the above method for making a Banach space out of C^∞ functions seems also to fail. An alternative possibility might be to try to use only certain C^∞ functions. After all, one would normally be concerned in some application only with having a "reasonable large collection" of solutions of one's equation (even though our remarks on page 76 show that there are even C^2 solutions). One might, therefore, proceed as follows. Fix positive numbers $a_0, a_1, \ldots$, and consider the collection of all C^∞ g such that $a_0\,\text{lub}\,|g| + a_1\,\text{lub}\,|g'| + \ldots$ is finite (this sum being the norm).

In this case, there is a constraint on what we can choose for the a_n: It must be true that the derivative of an admissible function (by the criterion above) is also admissible. That is to say, the a_n must be such that, if $a_0\text{lub}\,|g| = a_1\text{lub}\,|g'| + \ldots$ is finite, then $a_0\text{lub}\,|g'| + a_1\text{lub}\,|g''| + \ldots$ is also finite. But this in turn will be true if and only if the sequence of numbers whose n^{th} term is a_n/a_{n+1} is bounded. But now we are restricting ourselves essentially to analytic functions (for, given a function admissible by our criterion, expand it in Taylor series with remainder, and using boundedness a_n/a_{n+1}, show that the remainder term goes to zero as the Taylor series is made longer). It appears, therefore, that one can indeed find a suitable class of analytic functions. Unfortunately, restriction to

analytic functions is too severe for, (e.g., the physical) applications of our original partial differential equations: One wants to admit at least more C^∞ functions than these.

Having now failed in several attempt to find a suitable class of functions which will form a framework for the geometrization of our partial differential equation, it may be worthwhile to just list the properties we are looking for. We seek a collection E of mappings from R to R such that:

1. Linear combinations of functions in E are in E.

2. There is a norm on E which makes it a Banach space.

3. The derivative of any function in E is in E, and "take the derivative " is a continuous (linear) mapping from E to E.

4. The set E includes a "reasonable number" of functions. Our conclusion is, then, that we have been unsuccessful in finding such a collection.

It seems reasonable, next, to see if one can find some theorem which will crystallize our failure. We give one example of a result along these lines. There exists no collection E of mappings from the open interval $(0, 1)$ to R such that i) E is a Banach space, ii) each function in E is differentiable, its derivative is in E, and "take the derivative" is a continuous mapping from E to E, iii) there is at least one function in E which is constant on $(1/2, 1)$ and which is not constant on $(0, 1)$, and iv) if g is in E and $0 \leq a \leq 1$, then the function g_a with action $g_a(x) = g(2x - a)$ is in E; furthermore, for fixed g, this mapping from $(0, 1)$ to E is C^1. The proof is quite easy. Since "take the derivative" is a continuous linear mapping from E to E, we obtain a corresponding vector field ξ on the manifold E. This vector field has unique integral curves. Let γ be such an integral curve, so, for each $t, \gamma(t) = g_t$ is an element of E. Then the statement that this is an integral curve is: $d/dtg_t = g'_t$. Let g be the function in iii) above. Then the γ with $\gamma(t)(x) = g(2x-1+2t)$ is, by iv), an integral curve of our vector field. Its initial point, by iii), is a constant function. But an alternative integral curve is the constant one, $\tilde{\gamma}(t)(x) = g(2x - 1)$. By iii), these two integral curves are different. This contradiction establishes that no such collection of functions exists.

It seems, therefore, that either one must abandon the possibility of expressing partial differential equations geometrically on manifolds, or one must severely modify the rules. The first alternative would, in my view, be a severe blow to the potential applications of this subject. As to the second alternative, one could imagine at least two possibilities. On the one hand, one might look for something else (e.g., not even "functions") which one could still regard as "solutions" of our partial differential equation, and which could me made into Banach space. [The situation here is a bit reminiscent of that in Fourier analysis. The subject was a terrible mess until it was realized that one must work with the "right" set of functions – there, L^2.] One possibility which comes to mind, but which does not perhaps look too promising, is distributions. Perhaps a less severe notion of "derivative" needs to be incorporated. On the other hand, one might attempt to redo the subject of manifolds, basing them instead on topological vector spaces with less structure than Banach spaces. One possibility would be Frechet spaces (in which, essentially, the norm is replaced by a metric, but completeness is maintained). For example, the C^∞ functions with the "d" of page 78 form a Frechet space. The problem here is that it is apparently false in Frechet spaces that vector fields have unique integral curves (for the result of the previous paragraph fails if i) is replaced by "E is a Frechet space"). Perhaps there is some other kind of space in which one can work, such that reasonable classes of functions can be made into such a space, and such that, in this space, vector fields do have unique integral curves. The result of the previous paragraph, again, puts severe limitations on the possibilities.

The fact that the program is so successful and elegant for ordinary differential equations, and so appealing for partial, suggests that there should be some way to make it work.

26. Riemannian Geometry

In this section, we discuss briefly which of the various notions from Riemannian geometry in finite dimensions can be carried over to infinite.

Fix a Banach space E. A *metric* (on E – not to be confused with "metric" of "metric space") is an element g of the tensor space $\mathcal{L}(E, E; \mathbb{R})$ satisfying the following three conditions:

1. The tensor g is positive-definite. That is, for any x, y in E, $g(x, y) = g(y, x)$.

2. The tensor g is positive-definite. That is, for any nonzero x in E, $g(x, x) > 0$

3. The tensor g is invertible. First note that the tensor space $\mathcal{L}(E, E; \mathbb{R}$ is naturally isomorphic with $\mathcal{L}(E; \mathcal{L}(E; \mathbb{R})$; hence, we may (and often shall) regard g as an element of the latter. We require that there exist a g' in $\mathcal{L}(\mathcal{L}(E; \mathbb{R}); E)$ such that $g' \cdot g$ and $g \cdot g'$ are the identities on E and $\mathcal{L}(E; R)$, respectively.

Example. Let E be the Banach space of sequences of real numbers, $(r_1, , r_2, \ldots)$, the sum of the squares of whose entries converges. Let $a_1, a_2, \ldots)$ be a sequence of real numbers, and let g have the following action: For $x = (r_1, \ldots)$ and $y = (s_1, \ldots)$ in E, $g(x, y) = a_1\, r_1\, s_1 + a_2\, r_2\, s_2 + \ldots$. When is this g a metric? First note that, in order that g be well-defined, i.e., in order that the sum on the right converge for every x and y in E, it is necessary and sufficient that the $|a_i|$ be bounded. Then g is automatically symmetric. For positive-definiteness, it is necessary and sufficient that each a_i be positive. For invertibility, first identity $\mathcal{L}(E; \mathbb{R})$ with sequences of reals, so the action of $\mu = (q_1, q_2, \ldots)$ in $\mathcal{L}(E; \mathbb{R})$ on $x = (r_1, r_2, \ldots)$ in E is the real number $\mu(x) = q_1\, r_1 + q_2\, r_2 + \ldots$. Then, regarding g as in $\mathcal{L}(E; \mathcal{L}(E; \mathbb{R}))$, and applying to x, we have $g(x) = (a_1\, r_1, a_2\, r_2, \ldots)$. Clearly, then, its inverse g' must have action $g'(\mu) = (q_1/a_1, q_2/a_2, \ldots)$. But this will only be a bounded linear mapping, indeed, will only be well-defined, provided the $1/a_i$ are bounded, i.e., provided the a_i are bounded away from zero. This, then, is the condition for invertibility.

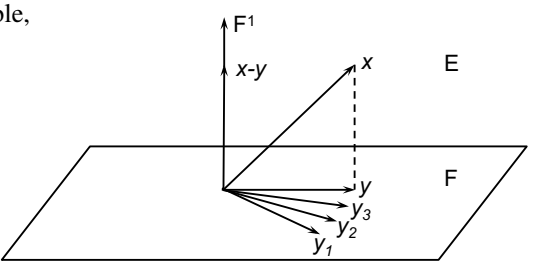
In the finite-dimensional case, the third condition above follows from the first two (i.e., in matrix terms, every finite-dimensional, symmetric, positive-definite matrix is invertible). As the example above, shows, this is not so in infinite dimensions. By the third condition, a metric establishes an isomorphism between E and its dual, $\mathcal{L}(E; \mathbb{R})$. Indeed, there is a kind of symmetry between E and its dual in this set-up. For example, g' can be regarded as an element of $\mathcal{L}(\mathcal{L}(E; \mathbb{R}, \mathcal{L}(E; \mathbb{R}); \mathbb{R})$ (namely, sending (μ, ν), both in $\mathcal{L}(E; \mathbb{R})$ to $\mu(g'(\nu))$. Then this element of $\mathcal{L}(\mathcal{L}(E; \mathbb{R}), \mathcal{L}(E; \mathbb{R}); \mathbb{R})$ is also symmetric, positive-definite, and invertible.

Let E be a Banach space with metric g. For x in E, set $\{x\} = (g(x, x))^{1/2}$. Then this "{ }" is actually a norm on E. [Proof: All the properties are immediate except the triangle inequality. Fix x and y in E. Then, by condition 2, we have that $g(x+ay, x+ay) = g(x, x)+ ag(y, x) + ag(x, y) + a^2 g(y, y) = g(x, x) + 2ag(x, y) + a^2 g(y, y)$ must be non-negative for every number a, where we used the first condition in the second step. But this quadratic polynomial in a can be non-negative only if $(g(x, y))^2 \leq g(x, x)g(y, y)$. Now set $a = 1$ in the above, to obtain $g(x + y, x + y) \leq g(x, x) + 2(g(x, x)g(y, y))^{1/2} + g(y, y)$, where we used our inequality. Taking the square root of each side, we obtain the triangle inequality.] We claim that in fact this norm "{ }' is equivalent to the norm "| |" that comes with the Banach space E. Indeed, we have $\{x\} = (g(x, x))^{1/2} \leq (|g|\,|x|\,|x|)^{1/2} = |g|^{1/2}|x|$. Suppose next that the reverse inequality, $|x| \leq a\{x\}$, for some a, were false. Then we could find a sequence of vectors in $E, x_1, x_2, \ldots$ such that $|x_i| = 1$, and such that $g(x_i, x_i)$ approaches zero. Now, $|x_i| = |g'(g(x_i))| \leq |g'|\,|g(x_i)|$, whence the $|g(x_i)|$, whence the $|g(x_i)|$ must be bounded away from zero. By definition of the norm of an element (such as $g(x_i)$) of $\mathcal{L}; \mathbb{R})$, therefore, there exist vectors $y_1, y_2, \ldots$ in E such that $|y_i| = 1$ and $|g(x_i)(y_i)|$ is bounded away from zero. But $|g(x_i)(y_i)|^2 = |g(x_i, y_i)|^2 \leq g(x_i, x_i)g(y_i, y_i) \leq |g|\,|y_i|^2 g(x_i, x_i) = |g|g(x_i, x_i)$, while

the last expression approaches zero as i approaches infinity. This contradiction establishes that $|x| \leq a\{x\}$ for some a. We conclude that our two norms are equivalent.

A Banach space possessing an equivalent norm which arises from a metric g as above is said to be *Hilbertable*. We conclude, therefore, that a Banach space possess a metric if and only if it is Hilbertable. In this case, we might just as well use the norm "{ }", which comes from a metric, rather than the original norm "| |". A Banach space, endowed with such a norm, is called a (real) *Hilbert space*.

Which Banach spaces are Hilbertable, i.e., which possess metrics? The answer is: Very few indeed. We next amplify this remark, and also give an interesting consequence of the existence of a metric.



Let E be a Banach space with metric g. Two vectors, x and y, in E will be said to be *orthogonal* if $g(x, y) = 0$. Now let F be any subspace of E. Denote by $F^{\perp}$ ("F perp") the set of all vectors in E which are orthogonal to every vector in F. We claim, first, that this $F^{\perp}$ is a subspace of E. Indeed, $F^{\perp}$ is clearly a vector subspace, for any linear combination of vectors, each orthogonal to every vector in F, is itself orthogonal to every vector in F, by linearity of g. Furthermore, $F^{\perp}$ is closed. [Proof: Let $x_1, x_2, \ldots$ be in $F^{\perp}$, and let these vectors converge to x. Then, for y in F, we have, since $g(x_i, y) = 0$ and since $\lim g(x_i, y) = g(\lim \mathrm{x_i}, \mathrm{y})$, that $g(x, y) = 0$. Thus, x is also in $F^{\perp}$.] We next claim that, not only is this $F^{\perp}$ a subspace, but that it is in fact complementary to F. Fix vector x in E, and set $a = \mathrm{glb}\,|\mathrm{x} - \mathrm{y}|$, where the "glb" is over all y in F, and where we use the norm which comes from g. Let $y_1, y_2, \ldots$ be vectors in F such that $\lim |x - y_i| = a$. We have the equality $|y_i - y_j|^2 = 2|x - y_i|^2 + 2|x - y_j|^2 - 4|x - (1/2)(y_i + y_j)|^2$, which follows immediately by expanding each side using linearity of g. For i and j sufficiently large, each of the first two terms can be made as close as we wish to $2a^2$. The last term, however, is greater than or equal to $4a^2$, by definition of a. Thus, for i and j sufficiently large, the right side is as small as we wish. We conclude: The y_i form a Cauchy sequence. Let y be the vector (necessarily in F, since F is closed) to which these y_i converge. Then we have $|x - y| = a$. We next claim that $x - y$ is in $F^{\perp}$. Indeed, for any z in F and any number b, we have $a^2 \leq |x - y + bz|^2 = |x - y|^2 + 2b\, g(x - y, z) + b^2|z|^2 = a^2 + 2b\, g(x - y, z) + b^2\, |z|^2$, where the first inequality follows from the definition of a and the fact that $-y + b\, z$ is in F. Subtracting "a^2" from each side, we obtain: $0 \leq 2b\, g(x - y, z) + b^2|z|^2$. But this quadratic-polynomial inequality in b can hold for all b (as it must) if and only if $g(x - y, z) = 0$. Thus, $x - y$ is in $F^{\perp}$. But now we have $x = y + (x - y)$, an expression for x as the sum of one vector in F and another in $F^{\perp}$. That is, F and $F^{\perp}$ are complementary.
[Its is almost immediate from this result that, e.g., $F^{\perp\perp} = F$.]

In particular, the above shows that, if a Banach space admits a metric, then every subspace of that Banach space split. Since we have seen examples of Banach spaces having subspaces which do not split, not every Banach space has a metric.

As usual, we now wish to pass from tensor spaces to fields. Let M be a C^p ($p \geq 1$) manifold based on Banach space E. A *metric* (field) on M is a C^{p-1} $\mathcal{L}(E, E; \mathbb{R}-$field which is symmetric, positive-definite, and invertible. Then, for example, its inverse g' is a $\mathcal{L}(\mathcal{L}(E; \mathbb{R}); E)-$field. Thus, a metric on M gives a sort of "local distance between nearby points of M" (regarding the "infinitesimal displacement between nearby points" as being represented by a tangent vector, and the norm of this vector, defined by g, as the "distance between these two points"). A metric, then, endows M with a "local geometry". There are also various senses in which one obtains a "global geometry". We give an example.

Let M be a C^p ($p \geq 1$) manifold based on Banach space E, and let g be a metric on M. Let γ be a curve on M, and let $a < b$ be two numbers in the interval on which γ is defined. Then, for each r in $[a, b]$, $\gamma(r)$ is a point of M, and the tangent vector to γ at

r, ξ_r, is a tangent vector at $\gamma(r)$. Further, $g(\gamma(r))$ is a $\mathcal{L}(E, E; \mathbb{R})$−tensor at $\gamma(r)$. Hence, $g(\gamma(r))(\xi_r, \xi_r)$ is a non-negative number. Set $L = \int_a^b (g(\gamma(r))(\xi_r, \xi_r))^{1/2} dr$, the definite integral of one real function of a real variable. This L is called the *length* of the curve γ from a to b. [Intuitively, one "sums the infinitesimal distances between successive points along the curve from $\gamma(a)$ to $\gamma(b)$".] This definition is a reasonable one, a remark we may illustrate by the following observation. Let us reparameterize our curve, i.e., choose a smooth monotonic function f of one real variable, and set $\hat{\gamma} = \gamma \cdot f$. Let $a = f(\hat{a})$ and $b = f(\hat{b})$, so $\hat{\gamma}(\hat{a}) = \gamma(a)$ and $\hat{\gamma}(\hat{b}) = \gamma(b)$. Then, we claim, the length of γ from a to b is the same as the length of $\hat{\gamma}$ from $\hat{a}$ to $\hat{b}$. Indeed, we have $\hat{\xi}_{\hat{r}} = f' \xi_r$ (where prime denotes $d/d\hat{r}$), whence $g(\hat{\gamma}(\hat{r}))(\hat{\xi}_{\hat{r}}, \hat{\xi}_{\hat{r}}) = (f')^2 g(\gamma(r))(\xi_r, \xi_r)$. Thus, the integrals defining the length are equal by change of the independent variable. Geometrically, this is of course what one would expect.

We can now use this notion of the length of a curve to define a "global distance function" on a manifold with metric. Let M, p, E, and g be as usual, and suppose further that M is connected, i.e., has the property that any two of its points may be joined by some curve. Given any two points, p and p', of M, set $d(p, p)$ the greatest lower bound of the length of curves joining p and p'. This "d", we claim, is a metric. [Proof: That $d(p, p') \geq 0$, and that $d(p, p') = d(p', p)$, are immediate. That $d(p, p') + d(p', p'') \geq d(p, p'')$ is also clear, since one curve from p to p'' is obtained by first tracing a curve from p to p', and then from p' to p''. We have only to show, therefore, that $d(p, p') = 0$ implies $p = p'$. Introducing a chart containing p, it suffices to do this within the Banach space E. Since, furthermore, the metric field is continuous, it suffices to consider the case when the metric is constant (for one can always find a "lower bound metric" in some neighborhood of p). Thus, we must show: For E a Banach space, g a metric on E, and p and p' distinct points of E, $d(p, p') > 0$. Let f be the function on E which assigns to x in E $f(x) = (g(x - p, x - p))^{1/2}$. Then, along any curve from p, f increases more slowly than the length of the curve increases (since the rate of change of f is measured by the norm of the component of the tangent vector in the direction away from p while the rate of change of length is measured by the norm of the tangent vector). Hence, $d(p, p') \geq (g(p' - p, p' - p))^{1/2}$.] Thus, a manifold-with-metric has the structure of a metric space.

The discussion above is a brief summary of "the geometry of manifold-with-metrics". We conclude this section with a discussion of metrics from a more algebraic point of view.

One of the most important and useful properties of a metric (field) is that it leads to a unique derivative operator. More precisely, we have: Let M be a C^p ($p \geq 2$) manifold based on Banach space E, with C^{p-1} metric g. Then there is one and only one derivative operator ∇ on M satisfying $\nabla g = 0$. The proof is computational and completely straightforward: One writes down the answer. Let U, ψ be a chart, and $\hat{g}$ the component of g. For u, v, and w in E, consider $1/2(D\hat{g}(u)(v, w) + D\hat{g}(v)(u, w) - D\hat{g}(w)(u, v))$. Fixing u and v, this assigns, to each w in E, a real number, i.e., is an element of $\mathcal{L}(E; \mathbb{R})$. Applying $\hat{g}'$ to this mapping, therefore, we obtain an element of E, an element which depends, of course, on our choices of u and v. Write this element $\Gamma(u, v)$, so Γ is an element of $\mathcal{L}(E, E; E)$. We now claim: This Γ is a connection. [Sketch of proof: Change the chart. We know the change in the component of g, and hence can compute, from the formula above, the change in Γ. The result is precisely the formula for the change in the component of a connection under chart-change.] Thus, we obtain a derivative operator on M. This ∇ further satisfies $\nabla g = 0$, for this equation, written in terms of a chart, is precisely the formula above used to define ∇.

Thus, once one has a manifold with metric, one automatically has a derivative operator, and hence write differential equations on tensor fields on M.

We remark that, just as in the finite-dimensional case, one can define the curvature tensor and obtain its symmetries (although one has neither the Ricci tensor, the conformal tensor, nor the scalar curvature, since these require a "contraction" which is not available in infinite dimensions), geodesics, etc.

Appendix

Problems

1. Find an example of Banach space E, vector subspace A of vector space E, such that A is finite co-dimension but not closed.

2. Let E be a Banach space. Show that every finite-dimensional subspace of vector space E is closed.

3. Find an example of a vector space, and two norms thereon each of which makes this vector space a Banach space, such that these norms are not equivalent.

4. Show that a subspace of a Banach space containing a ball is the entire Banach space.

5. Let E be a Banach space. Find all subspaces of E having a unique complementary subspace.

6. By a basis for Banach space E, we mean a sequence $x_1, x_2, \ldots$ of vectors in E such that, given any vector x in E, there is one and only one sequence $a_1, a_2, \ldots$ of real numbers such that the sequence in E whose n^{th} term is $y_n = a_1\, x_1 + \ldots + a_n\, x_n$ converges to x. [That is, every element of E can be written as an "infinite linear combination" of the x_i.] Show that our examples of Banach spaces have bases. Find an example of Banach space without basis.

7. Find a vector space having no norm which makes it a Banach space.

8. Let E and F be Banach spaces, and let T be a linear mapping (of vector spaces) from E to F. Denote by A the subset of $E \times F$ consisting of elements of the form $(x, T(x))$, with x in E. Prove that A is a vector subspace. Show that T is bounded if and only if A is closed in $E \times F$. [Hint: Open mapping theorem.]

9. Let E consist of all sequences, $(r_1, r_2, \ldots)$ of real numbers with $|r_2 - r_1| + |r_3 - r_2| + \ldots$ finite. Let the norm of such a sequence be this sum, plus $|r_1|$. Prove that this E is a Banach space.

10. Show that a Banach space is finite-dimensional if and only if every subspace is closed.

11. Define the product of an infinite number of Banach spaces.

12. Give an example to show that the open mapping theorem is false for normed vector spaces (i.e., without completeness).

13. Prove that every subspace of the Banach space E^2 (page 12) splits.

14. Prove that, if every subspace of E splits and every subspace of F splits, then every subspace of $E \times F$ splits.

Robert Geroch

General Relativity

1972 Lecture Notes

Robert Geroch
Enrico Fermi Institute
University of Chicago

© Robert Geroch
All rights reserved. Published 2013

ISBN: 978-0-9879871-7-4 (softcover)
ISBN: 978-0-9879871-8-1 (ebook)

Minkowski Institute Press
Montreal, Quebec, Canada
http://minkowskiinstitute.org/mip/

For information on all Minkowski Institute Press publications visit our website at
http://minkowskiinstitute.org/mip/books/

Preface

This publication of Robert Geroch's course notes on general relativity is the first book in the new *Lecture Notes Series* of the *Minkowski Institute Press*. The idea of this series is to extend the life in space and time of valuable course notes in order that they continue to serve their noble purpose by bringing enlightenment to the present and future generations.

Geroch's lecture notes on general relativity are unique in three main respects. First, the physics of general relativity and the mathematics, which describes it, are masterfully intertwined in such a way that both reinforce each other to facilitate the understanding of the most abstract and subtle issues. Second, the physical phenomena are first properly explained in terms of spacetime and then it is shown how they can be "decomposed" into familiar quantities, expressed in terms of space and time, which are measured by an observer. Third, Geroch's successful pedagogical approach to teaching theoretical physics through visualization of even the most abstract concepts is fully applied in his lecture on general relativity by the use of around a hundred figures.

Although the book contains lecture notes written in 1972, it is (and will remain) an excellent introduction to general relativity, which covers its physical foundations, its mathematical formalism, the classical tests of its predictions, its application to cosmology, a number of specific and important issues (such as the initial value formulation of general relativity, signal propagation, time orientation, causality violation, singularity theorems, conformal transformations, and asymptotic structure of spacetime), and the early approaches to quantization of the gravitational field.

Montreal, February 2013

Vesselin Petkov
Minkowski Institute Press

Contents

1 Introduction **1**

2 Events; Space-Time **3**

3 The Light-Cone Structure **7**

4 Clock Rates **9**

5 The Interval **11**

6 The Metric **15**

7 The Geometry of World-Lines **17**

8 Acceleration **21**

9 Acceleration and World-line Curvature **23**

10 Rotation **25**

11 Dust **29**

12 Perfect fluids **33**

13 Electromagnetic Fields: Decomposition by an Observer **35**

14 Maxwell's equation **37**

15 Stress Energy of the Electromagnetic Field **39**

16 Stress-Energy: A Summary **41**

17 The Optical Limit **43**

18 Geodesic Deviation **47**

19 Einstein's Equation **51**

20 Einstein's Equation: General Remarks **53**

21 The Friedmann Solution **55**

22 The Friedmann Solution: Continued **61**

23 Friedmann Solutions: The Observations **63**

24 Symmetries **69**

25 The Schwarzschild Solution **75**

26 The Schwarzschild Solution: Observations **81**

27 The Initial-Value Formulation **87**

28 Signal Propagation **93**

29 Time-Orientation **95**

30 Causality Violation **99**

31 An Implication of Absence of Causality Violation **101**

32 The Domain of Influence **103**

33 The Domain of Dependence **105**

34 Singularity Theorems **109**

35 Conformal Transformation **115**

36 Asymptotic Structure: Introduction **119**

37 Asymptotic Structure at Spatial Infinity **123**

38 Asymptotic Structure at Null Infinity **127**

39 Linearization **129**

40 Quantization of Gravitational Field: Introduction **133**

41 Linearized Approach to Quantization **135**

42 Canonical Approach to Quantization **137**

1. Introduction

The general theory of relativity is the most attractive classical (i.e., non-quantum) theory of the gravitational field available today. It represents a "generalization" of Newtonian gravitation which is consistent with the principles of special relativity.

Recall Newtonian gravitation. The gravitational potential is determined from the distribution of matter by the equation

$$\nabla^2\varphi = 4\pi G\rho \tag{1}$$

where G is the gravitational constant and ρ is the density of matter. This potential then produces a force, $F = -m\,\nabla\varphi$, on a particle of mass m. Thus, a distribution of matter produces a potential φ which, in turn, affects particles: in this way bodies interact gravitationally. What is wrong with this theory? The problem is that the distribution of matter at any one instant of time determines the potential at that time. This action is instantaneous. In other words, a particular "time" is singled out. Such a situation is simply not permitted by the principles of special relativity.

Thus, the phenomenon of gravitation certainly exists, and special relativity is about as well-established as a physical theory can be. One is naturally led to search for a relativistic theory of gravitation.

One can think of special relativity as the "limit" of general relativity when gravitation is unimportant, and of Newtonian gravitation as the limit when speeds are much less than that of light (so the Newtonian picture of space and time is a good approximation to the special relativistic picture). (Thus, the statement one occasionally hears that general relativity must replace special relativity when "accelerated frames" are considered is false. Special relativity must be abandoned when gravitation is important.)

It turns out that the general theory of relativity (i.e., this incorporation of gravitation into special relativity) touches on the structure of space and time. Why should this be true for gravitation (and not, say, for electromagnetism)? It is because gravitational phenomenon has about it an "all-inclusiveness" which is apparently not possessed by other forces in Nature (e.g., electromagnetic). Thus, particles can be positively charged, negatively charged, or uncharged: particles can react to electromagnetic fields in different ways. But all particles seem to react to gravitation in the same way. (Quantum theory is a good example of a physical theory with a similar "all-inclusiveness." The principles of quantum mechanics are not to be applied just to certain special systems. Quantum theory makes a much stronger claim. All physical phenomena are to be governed by its principles.)

This feature of gravitation is made more explicit by considering the famous elevator. People inside an elevator falling freely in a uniform gravitational field cannot distinguish their situation from that of an elevator at rest in no gravitational field. The uniform gravitational field affects everything – so by no physical experiment inside the elevator can these two situations be distinguished. (Like many important physical principles, this, the equivalence principle, becomes a tautology in the final theory. In general relativity, "uniform gravitational field" is a long way of saying "no gravitational field." That is, the existence of the former is simply not recognized. More on this later.)

The remarks above at least make it reasonable that gravitational phenomenon might occupy a special place in physics. They do not, however, make it clear

why this "special-ness" should manifest itself in statements about space and time. That this should be the case is suggested by the remarks below.

In order to do Euclidean geometry, one needs straight lines. I draw a line in the room. How do you determine whether or not it is straight? One might think of putting the edge of a meter stick against it. But, of course, that doesn't work unless one has a way to tell which meter sticks have straight edges. One might stretch a string between two points on the line, and see whether the string lies along the line. But a gravitational field will bend the string in a way over which we do not exert control. (The complete explanation here is more subtle. We shall return to this.) Finally, one might "sight" along the line to see if it is straight. Here, the implicit assumption is that light travels along straight lines. But it is known that a gravitational field affects the path of a light ray, for light from distant stars is observed to be bent of passing near the surface of the sun. Similar attempts fail for similar reasons. In short, since gravitation affects all physical phenomenon (all-inclusiveness), and since such phenomenon must be used, in particular, to make the measurements which determine the geometry of space (and time), it is perhaps not surprising that gravitation should be tied in with the structure of space and time.

2. Events; Space-Time

Recall the Newtonian picture of physics. One has some physical system he wishes to describe. At any one instant of time, that system is in one of a certain collection of possible states. As time goes on, the system evolves, i.e., it passes through a sequence of states. This general approach to describing a physical system is so natural that it is not always mentioned explicitly. We mention it because the description of things in relativistic physics is quite different. Time loses its special role. Instead, one adopts a description in which the past, present, and future of the system is all incorporated into one picture. This idea is reflected in the following.

A fundamental notion of general relativity is that of an *event*. An event is an occurrence (in the physical world) having no extension in either space or time. For example, the snapping of one's fingers, or the explosion of a firecracker, would represent an event.

The collection of all possible events – those which have happened, are happening, or ever will happen – will be denoted by M, and called *space-time*. This space-time is the arena in which all the physics takes place. An event is sometimes called a *point* of M.

The space-time of M of general relativity has more structure than just that of a set. We introduce part of this structure now. Suppose we consider a certain region of M (e.g., the collection of events which occur within a given room within of given interval of time). We wish to label points of this region by numbers. This could be done, for example, as follows. Let four pilots fly airplanes around the room. Each pilot carries with him a clock. (These clocks do not have to run at a uniform rate, or be synchronized.) Let an event which takes place in our region be the explosion of a firecracker. Each pilot records the time (as read from his clock) when he sees the flash. Thus, with this event, we associate the four numbers obtained by the pilots. These four numbers are called the *coordinates* of the event. Evidently, no two distinct events have exactly the same coordinates, and any set of coordinates (in the appropriate intervals) defines an event.

Evidently, there are many different systems of coordinates. The pilots can fly around in arbitrary ways, using clocks that run at arbitrary rates. Furthermore, there are other physical ways of obtaining systems of coordinates. One could, for example, fill the room, from floor to ceiling, with people during the time interval. Each person could be assigned three numbers which give his location in the room, and each person could carry a watch. Then, if an event takes place in the region of space-time, the person located at the occurrence of that event could record his watch's reading when the event takes place, and the three numbers which give his location. In that way, again, events are labeled by four real numbers.

Now suppose we select a region of space-time, and have two systems of coordinates, x^1, x^2, x^3, x^4 and y^1, y^2, y^3, y^4, in this region. That is to say, each event in our region is labeled by four numbers in two different ways. Hence (since values for the y-coordinates determine an event in the region, which, in turn, determines values for the x-coordinates), we obtain four functions of four variables, $x^1(y^1, y^2, y^3, y^4)$, $x^2(y^1, y^2, y^3, y^4)$, $x^3(y^1, y^2, y^3, y^4)$, $x^4(y^1, y^2, y^3, y^4)$. We assume that these functions are smooth (C^∞). (This is reasonable physically if the coordinate systems are set up with sufficient care. Thus, the pilots would have to move the controls of their airplanes smoothly, and their clocks would run smoothly, etc.) More generally, if the x's are an admissible system of coordinates, and if the y's are smooth functions of the x's, with the inverse functions (the x's as functions of

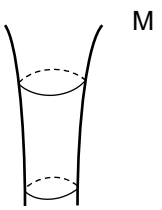
the y's) also smooth, we permit the y's as a system of coordinates.

Thus, on the space-time M we have a large number of admissible coordinate systems, any two smoothly related to each other, and with any system smoothly related to the admissible ones also admissible. But this is what the mathematical object called a four-dimensional manifold is. Thus, *the set M of events has the structure of a four-dimensional manifold.* (Such statements, which relate the physics to its mathematical description, are never proven. All one can do is make the mathematical description appear appropriate. Nor is it claimed that one can see, by abstract reasoning, that such a description will be useful. The foundations of a theory can be motivated, but its testing can only occur after the structure of the theory has been set down in detail.)

We wish to adopt the following point of view: every statement which is made about physical phenomenon is to be expressed in terms of the space-time manifold M. This is somewhat different point of view from that of Newtonian physics. The remarks bellow are intended to make the new attitude seems more natural.

How would one describe a particle in terms of space-time? Consider the set of all events which occur at the particle. That is, we consider events (snapping of fingers) such that, at the instant the fingers are snapped, the particle is right there. Evidently, this set of events describes a curve in space-time. This curve, called the *world-line* of the particle, describes completely the entire history (and future) of the particle.

Similarly, a string is described by the collection of all events which occur on the string – a two-dimensional surface in space-time. A bedsheet defines a three-dimensional surface in space-time. Two particles which collide have the property that their world-lines intersect. The point of intersection is the event of the collision of the particles.



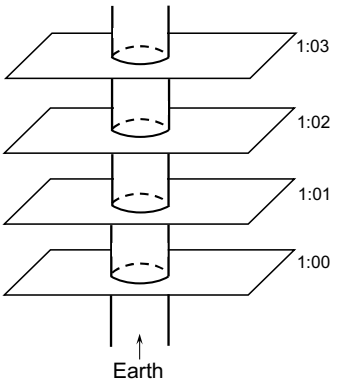
Note that there is no dynamics in space-time: nothing ever happens there. Space-time is an unchanging, once-and-for-all picture encompassing past, present, and future.

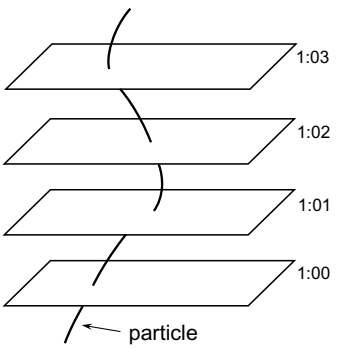
A star would be represented by a four-dimensional region of space-time. It would be tubular shaped, with (qualitatively) time going up the tube. If the star suddenly exploded, then the tube would "get wider" beyond a certain point. Collapse would be represented by a contracting of the tube.

There is, as we shall see later, no natural notion of "simultaneity" between events in space-time. But everyone near the Earth agrees when it is 1 : 00 Greenwich mean time. Suppose we fill the Earth (surface, interior, atmosphere) with people, all instructed to snap their fingers (mark the event) at 1 : 00 Greenwich mean time. These events would form a three-dimensional surface. If fingers were again snapped at 1 : 01, 1 : 02, etc,. we would obtain a family of such surfaces, filling the space-time. We can consider the first surface as representing "all of space at 1 : 00," the next as "all of space of 1 : 01", etc. We have now recovered simultaneity: two events are simultaneous if they occur on the same surface. We have also recovered dynamics: the situation at 1 : 00 is described by the space-time on the 1 : 00 surface, the situation at 1 : 01 by the 1 : 01 surface, etc. Things "change with time", for, as one moves from one surface to the next, one enters different regions of space-time, and so what is going on there might be different. A particle (i.e., a curve in space-time) would meet the successive surfaces at different points, and thus would be described as "moving in space as time evolves".

In short, we recover the Newtonian picture of physics by slicing the space-time with a family of three-dimensional surfaces. But, of course, there are many different ways to carry out such a slicing. There is no natural slicing, hence, no natural simultaneity, etc. All that is "natural" is the entire, four-dimensional space-time manifold. (There is nothing philosophical about this. It is just a different – and a rather attractive – way of describing what is going on.)

To summarize, an event is an occurrence having extension in neither space nor time. All the possible events (that have happened, are happening, or will happen) are collected together into space-time M. This M has the structure of a four-dimensional manifold. One adopts the point of view that everything is to be described in terms of structure on this manifold M.

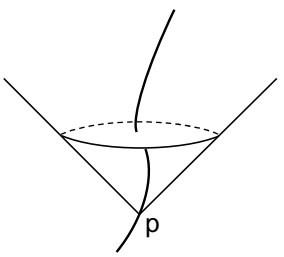




3. The Light-Cone Structure

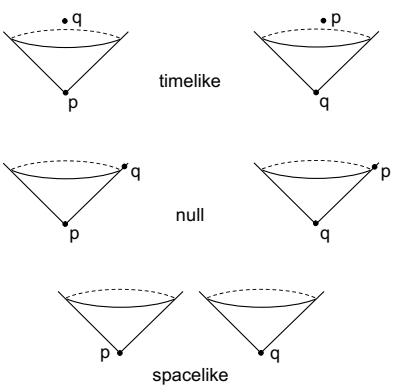
The next piece of structure on our space-time manifold M arises from the observation (in the real world) that no particles have been observed to move faster than light. thus the motion of light rays is of special significance.

The word "faster" refers to "speed", which refers to "distance traveled per unit time". Hence, the notion "faster" violates the spirit of space-time physics. We reformulate the observation in the first paragraph as follows. Suppose the explosion of a firecracker marks an event, so an expanding spherical pulse of light is sent out from it. Then any particle whose world-line passes through this event cannot get outside of the expanding sphere of light. (This is the statement that the particle's speed doesn't exceed that of light.) Let's try to reformulate this as a statement in space-time. Let the event (firecracker) be p. Then the light emerging from this event would form a cone with vertex p. (Slice the space-time with a family of three dimensional surfaces, and you will, indeed, see a "sphere of light expanding from the original event".) (In the figure, one dimension is, of necessity, suppressed.) Now we consider a particle (i.e., a curve) passing through p. The observation (about speeds not exceeding that of light) now reduces to the assertion that the world-line of the particles is inside the cone with vertex p.



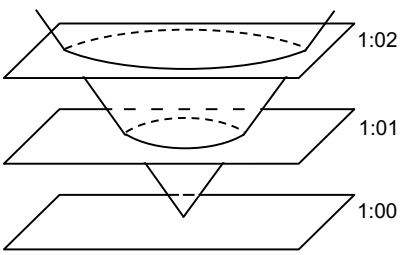
Thus, for each event of space-time there is a cone with vertex that event. This cone is the events which occur just as light from the original event passes by. The world-lines of particles which pass through the original event lie inside the cone. The cone is called the (future) *light cone*.

Now let p and q be two nearby events. (We need "nearby" because some notions, logically different, coincide for nearby events but do not coincide for more widely separated events. Of course, "nearby" means nearby in space-time.) If q lies inside the future light cone of p, we say that q is *timelike related* to p and to the *future* of p. If p lies inside the future light cone of q, we say that q is timelike related to p and to the *past* of p. If q lies on the future light cone of p (resp., p lies on the future light cone of q), we say that q is *null related* to p, and to the future (resp., past) of p. Finally, if neither event is either on or inside the future light cone of the other, we say the events are *spacelike related*. Evidently, a particle can meet first p and then q if and only if these events are timelike related, and p is to the past of q. A light ray can meet both events if and only if they are null related. Spacelike related events cannot be both met by any particle or light ray. (Note: A "light ray" corresponds to a flashlight which is on only momentarily, sending out a pulse of light. The light from a flashlight which is on continuously defines a two-dimensional surface in



space-time.)

It should now be clear that the three-dimensional surfaces introduced earlier (to obtain simultaneity) had the property that the future light cone from any event on any surface does not elsewhere intersect that surface. (Physically: light emitted at 1 : 00 cannot get to any other point by 1 : 00.)



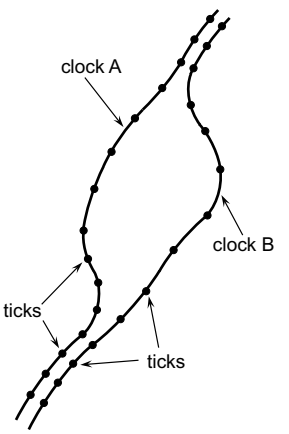
In special relativity, one can arrange matters, in an appropriate coordinate system, so that all the light cones look like ordinary cones, all with the same opening angle, and all parallel to each other. There is no reason, a-priori, to make such an assumption. In general relativity, no such assumption is made. Dropping it amounts to admitting curvature to space-time.

To summarize, one constructs a future light cone at each event using a light pulse emitted from that event. Particle world lines which pass through the event remain inside the light cone. The light cone leads to a distinction between timelike, spacelike, and null related events, and to a distinction between timelike, spacelike, and null related events, and to a distinction between past and future related events. (The past, future distinction is not available for spacelike related events.)

4. Clock Rates

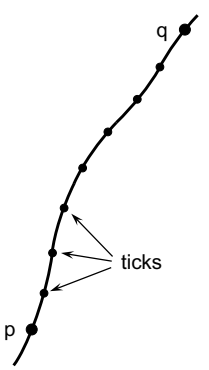
We shall shortly obtain a geometry (more precisely, a metric tensor field) on the space-time manifold M. This geometry is determined physically by measurements of time intervals between certain pairs of events. Finally, the possibility of making meaningful measurements of elapsed times rests on a certain assumption. We now introduce this assumption.

There exists clocks. We suppose that it is feasible to build clocks whose workings are not significantly affected by external influences (e.g., electromagnetic fields, accelerations, etc.). (This is a rather vague statement, whose meaning tends to change as our understanding of physics changes.) Now suppose we build two such clocks. Suppose, furthermore, that it is found that, when the clocks are held side by side (in space-time language, when the world-lines of the clocks are made to practically coincide over a certain stretch) their ticking rates coincide. (For example, this would, presumably, be true of two Cesium clocks.) Now the clocks are separated, and, for a while, traverse different paths in space-time. Finally, at some later time, the clocks are again held side by side. We assume that, *if the clocks initially ticked at the same rate, then when they are again brought side by side, they again tick at the same rate.* Note that we do not assume that the total elapsed times measured by the clock while they are separated are the same. In fact, general relativity predicts that this will not in general be the case. The assumption, essentially, is only that the ticking rate of an "isolated" clock along a certain stretch of its world-line depends only on the world-line in that region, not on what the world-line was before the region of interest. In other words, the ticking rate of a clock does not depend on what has happened to it in the past.



Why should one believe such an assumption? It would imply, for example, that two atomic clocks, one assembled in Chicago and the other in Bombay will, when they are brought together, tick at the same rate. Physically, one might expect this to be the case.

An "experimental test" of this assumption has been suggested by Wheeler. An electron has a mass, and hence a Compton wavelength, and hence a characteristic time (the light-travel time across a Compton wavelength). Hence, an electron is essentially a "clock". The electrons in an atom inside the Earth have, presumably, traversed different world-lines in space-time to get where they are. A "comparison of the ticking rate" is provided by the Pauli exclusion principle. It is because of this principle (i.e., because all electrons are identical) that after



the first few electrons in an atom have occupied the lowest states, the remaining electrons must occupy higher states. Now, a typical electron in an atom in the Earth has gone around the nucleus 10^{30} times during the age of the Earth. Hence, if mass differences between different electrons were more than about one part in 10^{30}, all the electrons would by now have dropped to the lowest level (because they would have "discovered" their mass differences, i.e., discovered that they do not have to obey the exclusion principle). But this collapse of atoms in the Earth has not taken place. Hence, our assumption has been "tested" to one part in 10^{30}.

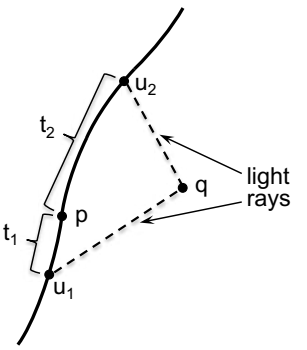
Now suppose we build one clock, and call the elapsed time between its ticks the second. Then any other clock can, by being brought side by side with this one, be recalibrated so that it ticks in seconds too. Once the second clock has been so calibrate, it can be carried around anywhere in space-time. We calibrate a large number of clocks in this way. Then, by our assumption, any two will, when brought side by side, tick at the same rate. In other words, if we have a world-line of a particle, and two events p and q on this world-line, then we can count the number of ticks (number of seconds) between the events p and q using a calibrated clock which traverses the world-line between p and q. Our assumption ensures that this number (of seconds) is independent of which calibrated clock we use. (Although it will in general depend on what the world-line is between p and q.)

Thus, *given a world-line between events p and q, it is meaningful to speak of the elapsed proper time (in seconds) along that world-line between p and q.*

5. The Interval

We can now speak of the elapsed proper time between two events along a world-line joining those events. This leads to the notion of the interval between two events.

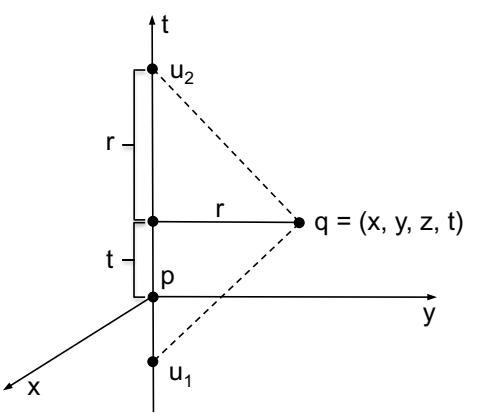
Let p and q be two nearby events in space-time. (More precisely, we require that the displacement of q from p be infinitesimal. That is, what follows is good only to first order in the displacement.) Arrange a world line of a particle to pass through event p. Now find a light ray which first intersects this world-line, and then event q. Find a second light ray which first meets q, and then intersects the world-line. Denote by u_1 and u_2 the two events at which these light rays intersect the world-line. Finally, let t_1 be the elapsed proper time along this world line between u_1 and p, and let t_2 be the elapsed proper time between p and u_2. (Thus, t_1 and t_2 are just a certain number of seconds.) We define the *squared interval* between event p and event q as the product of these times:



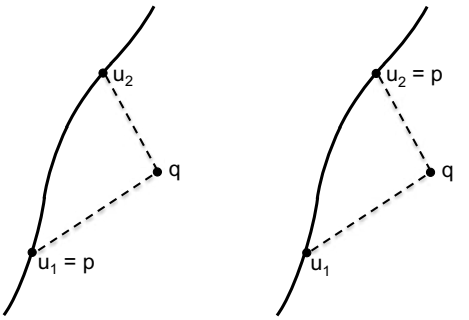
$$SI(p, q) = t_1 \, t_2 \tag{2}$$

We next require a further assumption: that the quantity $SI(p, q)$ depends only on the (nearby) events p and q, and not on the other details of the construction (whether the original world-line is passed through p or q; the choice of that world-line).

What are the grounds for this assumption? Let us see what the construction above yields in special relativity. Our space-time is Minkowski space, with the usual x, y, z, t coordinates. Let p be the event with coordinates $(0, 0, 0, 0)$, and q the event with coordinates (x, y, z, t). Let the world-line be the t-axis (the straight line $x = y = z = 0$). Then the parameter t along this world-line measures proper time (between events on this world line). Light rays are straight lines making a 45^o angle to the vertical (i.e., to the t-axis). That is, the light cones are all open 45^o from the vertical. In this case, the event u_1 has coordinates $(0, 0, 0, t - r)$, and u_2 coordinates $(0, 0, 0, t + r)$, where $r^2 = x^2 + y^2 + z^2$. Hence, $t_1 = r - t$, and $t_2 = r + t$. Thus, $SI(p, q) = t_1 \, t_2 = r^2 - t^2 = x^2 + y^2 + z^2 - t^2$. This quantity will be recognized as the usual squared interval between points in Minkowski space in special relativity. Clearly, the construction above yields the correct squared interval for any two events, and any straight world-line in special relativity.



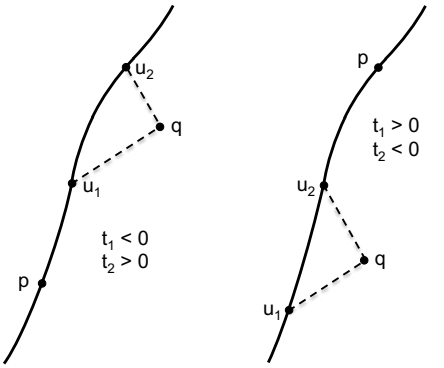
Of course, even in special relativity, the quantity $SI(p,q)$ will not be independent of the world-line if non-straight world-lines are admitted. It is clear intuitively, however, and easily checked, that independence of the world-line does hold in special relativity to first order in the displacement of q from p.



Thus, our assumption amounts, essentially, to the assumption that, *locally, space-time has the same structure as the space-time of special relativity.* More precisely, space-time is that of special relativity to first order in displacement. (Still more precisely, and in the language of general relativity, the displacement of p from q must be small compared with the radius of curvature of space-time.

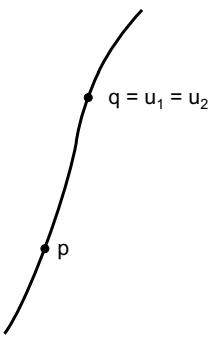
In the discussion above, q was spacelike related to p; and we had $SI(p,q)$ positive. Suppose that q were null related to p. Then $u_1 = p$ (if q is to the future of p), or $u_2 = p$ (if q is to the past of p). This is clear because "null related" means precisely that there is a light ray joining p and q. Hence, when q is null related to p, either $t_1 = 0$ or $t_2 = 0$ (the cases $u_1 = p$ and $u_2 = p$, respectively). In either case, $SI(p,q) = t_1\, t_2 = 0$

Suppose, finally, that q is timelike related to p. Then, if q is to the future of p, u_1 and u_2 are to the future of p, so t_1 is negative and t_2 positive. If q is to the past of p, then u_1 and u_2 are both to the past of p, so t_1 is positive and t_2 negative. Thus, in either case, $SI(p,q)$ is negative.

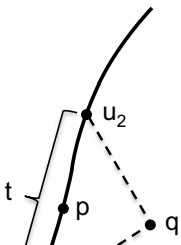


To summarize, $SI(p,q) = SI(q,p)$ *is positive if p and q are spacelike related, zero if p and q are null related, and negative if p and q are timelike related.*

Now suppose that p and q are timelike related (say with q to the future of p). Choose the world-line to pass through p and q. Then $u_1 = u_2 = q$ (since q is already on the world-line, the "light rays" from q to the world-line are very short). Hence, $-t_1 = t_2 =$ proper elapsed time between p and q along this world-line. (To first order, this quantity is independent of world-line). Hence, for p and q timelike related, $SI(p,q)$ *is minus the square of the proper elapsed time between p and q* (along any world-line joining the events).



Similarly, let p and q be spacelike related, and choose the world-line so $t_1 = t_2$. An observer on this world-line would interpret the situation as follows. He is bouncing a light ray off q, and measuring the round-trip light



travel time, $t = t_1 + t_2$. He would say that the "spatial distance" of q from p is $\frac{1}{2}t$ (where his unit of distance is the light-second, the distance light goes in a second). That is, *for p and q spacelike related, $SI(p, q)$ has the intuitive interpretation as the square of "distance" (measured in light-seconds) of q from p.*

In this sense, then $SI(p, q)$ is interpreted as representing geometrical information about spacetime.

6. The Metric

Imagine a very large book in which there is listed all possible pairs (p, q), of nearby events, and, for each pair, the number $SI(p, q)$. This book contains all the information about the geometry of space-time. One could in principle treat the geometry of space-time by constantly dragging out this book, but obviously, this would not be very convenient. One is thus let to attempt to describe the geometrical information in SI in a more convenient form.

Let p be an event of space-time, i.e., let p be a point of the manifold M. Recall that a nearby event, q, to p, is described by a contravariant vector, ξ^a, at p. We now introduce the assumption: there is a tensor $g_{ab} = g_{ba}$ at p such that

$$SI(p, q) = g_{ab}\xi^a\xi^b \tag{3}$$

for all "infinitesimally nearby" q. This assumption can be considered as a further extension of the idea that "locally, space-time has the structure of space-time in special relativity". In special relativity, expression (3) holds, where g_{ab} is the usual metric $(dx^2 + dy^2 + dz^2 - dt^2)$ of Minkowski space. We therefore assume that it holds at each point of space-time.

The above is repeated at each point of space-time. In this way, we obtain a symmetric tensor field g_{ab} on the manifold M. Since ξ^a connects infinitesimally nearby points which are null related precisely when $g_{ab}\xi^a\xi^b = 0$, and since the points null related to a point p are supposed to form a cone with vertex p, it is clear that this g_{ab} must have signature $(-, +, +, +)$. (Or, to state matters another way, we have special relativity locally.) We assume that g_{ab} is a smooth tensor field. (As a rule of thumb, whenever smoothness is meaningful for a physical quantity, there is no loss, physically, in assuming smoothness. Perhaps that's a reason why differential geometry – the study of smoothness – is so useful in physics.) From the statement of its signature, g_{ab} has an inverse, g^{ab}. Thus, g_{ab} is a metric tensor field of M.

To summarize, *the information in the $SI(p, q)$'s i.e., the geometry of space-time, is completely described by a metric tensor field g_{ab} on M of signature* $(-, +, +, +)$.

For obvious reasons, a vector ξ^a at a point is called *timelike*, *null*, or *spacelike* according as $g_{ab}\xi^a\xi^b$ is negative, zero, or positive, respectively.

We now have a four-dimensional manifold M with metric tensor field g_{ab} of signature $(-, +, +, +)$. All the technology of differential geometry can now be carried over to M. In particular, we have the notion of (indexed) tensor fields on M. Indices of such fields are raised and lowered with g_{ab} and its inverse. We have the (unique) derivative operator ∇_a on such tensor fields satisfying $\nabla_a g_{bc} = 0$. We have a Riemann tensor R_{abcd} satisfying the Bianchi identities, and having the algebraic symmetries $R_{abcd} = R_{[cd][ab]}, R_{[abc]d} = 0$. We construct from the Riemann tensor the Ricci tensor and scalar curvature.

The basic philosophy is that all physics is to be described in terms of space-time manifold M. We can now state this philosophy a bit more precisely: physics is to be described in terms of tensor fields on M. Among these fields, easily the most important is the metric, g_{ab}. There will soon be others.

We are now in a position to make more precise the distinction between special relativity and general relativity. *The framework we have set up above is called special relativity when the Riemann tensor vanishes* (everywhere), *and general relativity otherwise.* (It is easily checked that the vanishing of the Riemann tensor is

precisely the condition for the existence of a coordinate system in which the components of the metric take the usual special relativity form, diag $(-1, +1, +1, +1)$.) Suppose that one views Minkowski space (the space-time of special relativity) from a "uniformly accelerated frame", e.g., an accelerated elevator. No matter how the situation is "viewed", the Riemann tensor vanishes. In other words, special relativity is applicable.

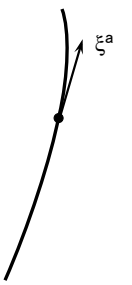
It is clear, from the way the metric was introduced, that it represents geometrical information about space-time. as we shall see shortly, the metric has a second interpretation: as the "potential of the gravitational field".

We conclude with an obvious remark. The metric is uniquely determined by certain physical measurements (of SI). Hence, any statement ones makes about the metric automatically has physical significance. Of course, some statements (about g_{ab}) may be easier to interpret than others, but all are in principle interpretable. One can freely do mathematics, introducing physical interpretations where they are appropriate and interesting.

7. The Geometry of World-Lines

We are now in a position to describe more precisely the world-lines of particles and light rays.

Consider a curve in space-time, i.e., a smooth mapping γ: $I \to M$, where I is some open interval of the real line. Then, for each number λ in I, the tangent vector to this curve, ξ^a, is a contravariant vector at the point $\gamma(\lambda)$ of M. This curve is said to be *timelike* if, for each value of λ, the tangent vector ξ^a is timelike; *null* if, for each λ, ξ^a is nonzero and null; and *spacelike* if, for each λ, ξ^a is spacelike. (Note: timelike and spacelike imply nonzero.)



The statement of Sect. 3 is now the following: *the world-line of a material particle is a timelike curve; the world-line of a light ray is a null curve.* (This whole business is a bit circular, as such arguments always are. We originally used the world-line structure to obtain the metric; now that structure is reexpressed in terms of the metric.)

Let us reparameterize our curve. Let $\lambda(\lambda')$ be a function of λ', so we replace the parameter λ by λ'. Then the corresponding tangent vectors, ξ^a and ξ'^a, are related by

$$\xi'^a = \xi^a \frac{d\lambda}{d\lambda'} \tag{4}$$

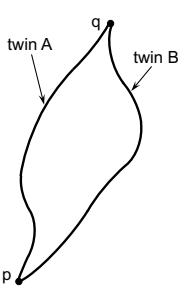
If p and q are two events on this world line, then, immediately from (4),

$$\int_p^q [-\xi'^a \xi'_a]^{1/2} d\lambda' = \int_p^q [-\xi^a \xi_a]^{1/2} d\lambda \tag{5}$$

Thus, the integral (5) depends only on the world-line and on the events p and q, not on the parameterization of the world-line. This number is called the *length* of the (timelike) world-line between p and q.

There is a simple physical interpretation for the length of a timelike world-line. Recall that $(-\xi^a \xi_a)^{\frac{1}{2}}$ is the proper elapsed time between the events whose infinitesimal displacement is represented by ξ^a. Hence, (5), the sum of these "infinitesimal elapsed times between successive points on the world-line" is just the elapsed proper time between p and q along this world-line.

Thus, we have that *the length of a timelike curve between two events is precisely the elapsed proper time* (in seconds) *between those events along the curve.* (This statement could have served as the physical definition of the metric.)



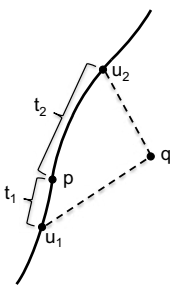
Suppose that two identical twins traverse world-lines which meet at event p, and then again at event q. In general, these two curves will not have the same length between p and q. Thus, in general, one twin will age more than the other between p and q. (This remark is sometimes called the "twin paradox". Note that the fact that different curves

between two points have different length doesn't require curvature. The difference in ageing is a special-relativistic effect.)

It is clear from Eqn. (4) that, given a timelike curve, one can always choose a parameter t for that curve such that its tangent vector is unit: $\xi^a\xi_a = -1$. (Such a t is uniquely determined up to the addition of a constant.) In this case, by (5), the proper time between two points on the curve is just the difference between the t-values of those points. One says that the curve is *parameterized by length*, or *parameterized by proper time*. The corresponding (unit) tangent vector to the curve is called the four-velocity (or just velocity) of the particle.

When we speak of an *observer*, we normally mean merely a timelike curve. (You and I, after all, are merely timelike curves in space-time.)

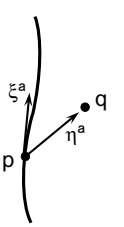
Consider an observer, and consider two events, p and q, with p on his world-line, and q infinitesimally displaced. Under what conditions would our observer say that that the events p and q occurred "simultaneously". One criterion our observer might use is the following. He might send out a light ray to intersect q, and receive a light ray from q (just as if he were going to determine $SI(p,q)$.) He might call the events "simultaneous" if $t_1 = t_2$. (Think about that a moment. If you sent a light ray to the sun at 1:52, and it reached the sun just as a solar flare began, and if you received a light ray from the flare at 2:08 (i.e., you saw the flare at 2:08), then you would say that the event "start of the flare" and "when my watch read 2:00" were simultaneous.) Let's express this condition $t_1 = t_2$ tensorially. Let λ^a be the vector representing the infinitesimal displacement between p and q, and let ξ^a be the four-velocity of the observer at p. Then $\lambda^a + t_1\xi^a$ is the displacement vector from u_1 to q, while $\lambda^a - t_2\xi^a$ is the displacement vector from u_2 to q. Since these displacements represent the paths of light rays, they must be null:



$$\begin{aligned} 0 &= (\lambda^a + t_1\xi^a)(\lambda_a + t_1\xi_a) = \lambda^a\lambda_a + 2t_1\lambda^a\xi_a - (t_1)^2 \\ 0 &= (\lambda^a - t_2\xi^a)(\lambda_a - t_2\xi_a) = \lambda^a\lambda_a - 2t_2\lambda^a\xi_a - (t_2)^2 \end{aligned} \tag{6}$$

These two equations can hold with $t_1 = t_2$ when and only when $\xi^a\lambda_a = 0$. In other words, *"pure spatial displacements" from p (according to this observer) are represented by vectors at p which are orthogonal to the four-velocity of the observer.*

Again, let p be an event on our observer's world-line, and let q by a nearby event. Denote by η^a the vector representing this displacement. Our observer would say that q occurred a certain "spatial distance" from p, and a certain "time earlier (or later)" than p. He would obtain such a description as follows. He first decomposes η^a into the sum of a multiple of his four-velocity and a vector orthogonal to his four-velocity:



$$\eta^a = -\xi^a(\eta^m\xi_m) + (\eta^a + \xi^a(\eta^m\xi_m)) \tag{7}$$

The first term represents what he would call the "temporal displacement" between p and q. Hence, he would say that the "elapsed time" between these events is:

$$t = [-(\xi^a(\eta^m\xi_m))(\xi_a(\eta^n\xi_n))]^{\frac{1}{2}} = -\eta^m\xi_m, \tag{8}$$

where the sign is chosen to give the elapsed time from p to q. The second term in (7) represents what he would call the "spatial displacement" between p and q. (This term is orthogonal to his four-velocity, and so represents, for him, a spatial displacement.) Hence, he would say that the "spatial distance" between p and q is

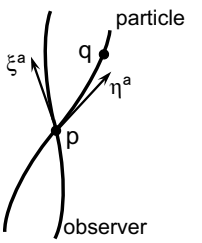
$$r = [(\eta^a + \xi^a(\eta^m\xi_m))(\eta_a + \xi_a(\eta^m\xi_m))]^{\frac{1}{2}} = [\eta^a\eta_a + (\xi^a\eta_a)^2]^{\frac{1}{2}} \tag{9}$$

Note, from (8) and (9), that

$$r^2 - t^2 = \eta^a\eta_a$$

i.e., the squared interval between p and q. A different observer (i.e., a different world-line through p) would assign different "spatial and temporal displacement" to p and q. But all would agree on the value of $r^2 - t^2$, the value of the squared interval.

In this sense, then, each observer has his own personal (local) decomposition of space-time into "space" and "time". In this sense, the "obvious" subjective difference between space and time, and their unification into space-time, can co-exist.



Quite generally, given a tensor at p, then any observer who passes through p can, by projecting the indices of that tensor parallel and perpendicular to his four-velocity, express that tensor in terms of tensor all of whose indices are perpendicular to his four-velocity.

We consider one further example of these ideas. Let an observer pass through event p, and also let a particle's world-line pass through p. Let the four-velocities be ξ^a and η^a, respectively. What is the apparent v "velocity of the particle" as seen by the observer? The particle leaves event p to go to event q displaced by η^a from p. Thus, the particle appears to go a spatial distance r (given by (9)) in time t (given by (8)). Hence, the apparent speed of the particle, as seen by our observer, is ($\eta^a\eta_a = -1$)

$$v = \frac{r}{t} = \frac{[(\xi^a\eta_a)^2 - 1]^{1/2}}{(-\xi^a\eta_a)} \tag{10}$$

Note that this speed is always less than one (i.e., less than the speed of light, since our units of speed are light-seconds/second).

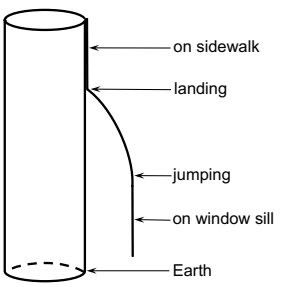
With each material particle there is associated a number m called the (rest) *mass* of the particle. The *four-momentum* of the particle is m times the four-velocity of the particle. An observer with four-velocity ξ^a computes $\xi^a p_a$, where p_a is the four-momentum, and calls this the apparent (i.e., to him) energy of the particle. Let's work out the apparent energy in terms of the apparent speed. We have, from (11)

$$\text{Energy} = -\xi^a p_a = -m\,\xi^a\eta_a = \frac{m}{\sqrt{1-v^2}} \tag{11}$$

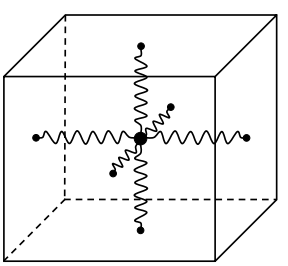
This formula is, of course, famous.

8. Acceleration

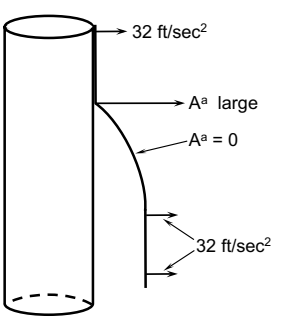
Some world-lines are more desirable (for an observer to follow) than others. Suppose you decide to jump off the fourth floor of Eckart. Your world-line would look like that in the figure (note the event "jumping", and the event "landing on the sidewalk"). Why is it that one suffers injuries at the event of meeting the sidewalk, rather than, say, at the event of jumping? One would normally say that he underwent a very large acceleration on meeting the sidewalk. In terms of the observer's world-line, this feature is reflected in the fact that his world-line has a "kink" at the event "meeting the sidewalk". Our purpose is to formulate the remarks above quantitatively.



We first remark that, at each event on his world-line, an observer can determine a quantity called his acceleration. Such a determination could, for example, be carried out as follows. Our observer constructs a cubic box inside of which there is a small mass, suspended by springs attached to the faces of the cube. (Such an instrument is called an accelerometer.) At each event on his world-line, our observer can measure the displacement of the small mass from its (normal) position at the center of the cube. This (spatial) displacement defines a vector (in space) for our observer. Hence, from the discussion of Sect. 7, we have a vector perpendicular to the four-velocity of the observer. (By convention, this vector, the acceleration, is defined as minus the displacement of the small mass from the center. Thus, if the cube is accelerated forward, the mass from the center. Thus, if the cube is accelerated forward, the mass would reside near the rear of the cube.)



To summarize, *at each event of his world-line an observer can, by means of a physical measurement, determine a vector A^a orthogonal to his four-velocity. This vector is called the acceleration of the observer (or, of the world-line).* (In other words, an observer's acceleration is an absolute, local quantity, measurable without reference to anything external.)



Let us now return to the observer who jumps off Eckart. While standing on the window sill before the jump, his acceleration is 32 feet/sec^2 (in magnitude), directed upward (the mass shifts to the lower half of the cube). After he jumps (i.e., in free fall), his acceleration is zero. This, of course, is just the reverse of the usual description (in which standing on the window sill is "not accelerating", and falling is "accelerating at 32 ft/sec^2 is response to the Earth's gravitational field"). The dif-

ference is that acceleration (as we have defined it) is not relative to the Earth's surface, or relative to anything. It is the reading from one's cube. One might say that we have "neglected to subtract off the effect of the Earth's gravitational field on the mass inside the cube". That is precisely what we have done... on purpose.

Let us suppose that there were an electric field present, and that our observer were charged. Then he would move in response to this electric field. The mass in the cube would record these accelerations provided that mass were uncharged. If, on the other hand, the mass were also charged (with the same charge/mass ratio as that of the observer holding the cube), then observer and mass in the cube would move together, and the cube would always measure zero acceleration. Why don't we use a charged mass in our cube? Because it is possible to "turn off" electromagnetic forces by making the mass neutral. Gravitational forces, on the other hand, cannot be "turned off". This is the all-inclusiveness of gravitational phenomena – the equivalence principle. Thus, the prescription is that we make the mass inside the cube insensitive to everything we can (i.e., insensitive to everything except gravitation).

We now adopt the attitude that, *whenever a particle's world-line experiences an acceleration, that acceleration must be attributable to some non-gravitational influence.* Thus, the observer standing on the window sill attributes his acceleration (32 ft/sec^2, upward) to the force the window sill is exerting on his feet. After jumping, the window sill no longer exerts that force, and, indeed, the observer measures (with his cube) zero acceleration.

Finally, consider the elevator. Observers inside the elevator cannot tell whether the elevator is sitting on the surface of the Earth, or the elevator is being accelerated through space (at 32 ft/sec^2) by a rope attached to it. But, from the present point of view, their description of the two situations are the same. In both cases, they would say that they are accelerated because of the force the floor of the elevator exerts on their feet. Thus, the equivalence principle (physical indistinguishability of a uniform gravitational field from acceleration) leads, in general relativity, to calling both phenomenon the same thing: acceleration.

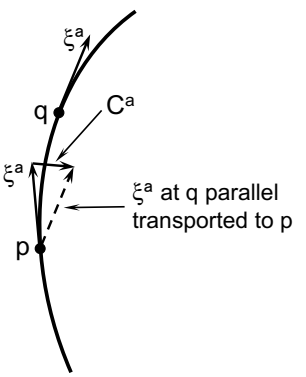
The discussion above is intended to bring out a very important point of view in general relativity. *An observer does not* (because he cannot, in principle) *attempt to decompose his acceleration into a "real acceleration" and an "apparent acceleration due to gravitational forces on the small mass in his cube". These two quantities* (which are distinct in Newtonian gravitation) *are combined into a single, unambiguous quantity – acceleration – measured by the displacement of the small mass inside the cube.*

Note: our units of length are light-seconds (distance light travels in a second). So, 1 ft = 10^{-9} light-seconds. So, acceleration has units light-seconds/second^2 = sec^{-1}. Thus, 32 ft/sec^2 = 3×10^{-8} sec^{-1}.

9. Acceleration and World-line Curvature

We have seen in Sect. 8 that, at each point of a world-line there is associated a vector A^a, the acceleration vector of the world-line, which is orthogonal to the four-velocity at that point. We now want to relate this (physically measured) quantity to some geometrical (mathematically expressed) property of the world-line.

Let us return again to the figure on p. 31. The observer in that figure suffers a very large acceleration at the event "landing on the sidewalk". Is there any way in which this physical fact appears in the "shape" of his world-line at that event? Of course there is: his world-line has a kink at that event. Thus, we expect that acceleration should somehow be related to the extent to which a world-line "bends".



There is a convenient measure of "bending". Let $\gamma\colon I \to M$ be a timelike curve, parameterized by length. Denote by ξ^a the corresponding tangent vector to this curve. The quantity

$$C^a = \xi^m \nabla_m \xi^a$$

is called the *curvature* of the world-line. Note that the curvature is essentially the derivative of the four-velocity in the direction of the four-velocity. This curvature has the expected geometrical interpretation: it is large for a world-line which "curves sharply", etc. Note also that the curvature of a curve vanishes if and only if that curve is geodesic. Note also that the curvature of a world-line is necessarily orthogonal to the tangent vector to that world-line:

$$\xi_a C^a = \xi_a(\xi^m \nabla_m \xi^a) = \frac{1}{2}\xi^m \nabla_m(\xi_a \xi^a) = \frac{1}{2}\xi^m \nabla_m(-1) = 0.$$

We now introduce the following important assumption: *the* (physically measured) *acceleration of a world-line is equal to the* (geometrical) *curvature of that world-line*, $C^a = A^a$. Note firstly that this identification is at least reasonable because both vectors are orthogonal to the four-velocity. Furthermore, it is suggested by the first figure of Sect. 8. Finally, the curvature of a world-line is the right sort of quantity to equate to the acceleration, for the former is the derivative of the four-velocity, and so is an "acceleration-like" quantity.

Although these remarks make the assumption plausible, they by no means establish it conclusively. The assumption related space and time measurements (the measurements which give squared intervals, hence the metric, which is needed to find the curvature) to acceleration measurements (by an observer's cube). I know of no "conclusive" argument for it. In fact, one can write down theories in which it does not hold. It is simply an assumption (a very strong assumption with little experimental foundation) of general relativity. Perhaps the brunt of

the assumption is its assertion that there is not some additional tensor field on M which combines with curvature of the world-line to produce the acceleration.

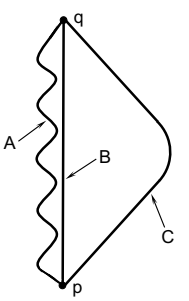
We shall hereafter normally use the word "acceleration" in place of "curvature of timelike world-line".

Let a particle of mass m travel on a certain timelike world-line. The quantity mA^a is called the *force* on the particle, F^a. (Newton's law: $F = ma$.) Noting that the four-momentum of the particle, p^a, is given by $m\xi^a$, where ξ^a is the four-velocity of the particle, we have $F^a = a\xi^b\nabla_b\xi^a = m^{-1}p^b\nabla_b p^a$. It is this force which must be ascribed to non-gravitational influences on the particle. We shall obtain examples of such forces (electromagnetic, pressure) shortly.

Note, in particular, that when there is no force on a particle (i.e., when we have a free particle), the world-line of that particle is a geodesic. *A free particle travels on a timelike geodesic.*

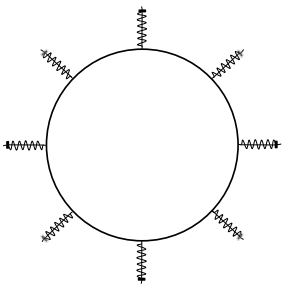
Recall the force law in Newtonian gravitation: $\vec{F} = -m\vec{\nabla}\varphi$. Our assumption ($A^a = C^a$) is essentially the relativistic version of this. Think of the metric g_{ab} as analogous to the "gravitational potential", φ. Than C^a involves (in terms of components) the connection, and hence first (coordinate) derivatives of the components of the metric. Thus, in each case, the force is expressed in terms of first derivatives of the potential. (The distinctive feature of general relativity is that g_{ab} has two interpretation: as a geometrical quantity and as a gravitational potential.)

We remark on an obvious consequence of the discussion above. It is common in physics to describe effects on particles in terms of a force field (e.g., electric field, Newtonian gravitational field, etc.). That is, one has a vector field F^a such that, at each point, the force on a particle at that point is F^a. No such force fields are allowed in general relativity. The reason is that the force F^a must be orthogonal to the four-velocity of the particle on which this force acts. If we were given F^a at the beginning, then we could always find some timelike world-line to which F^a were not orthogonal.



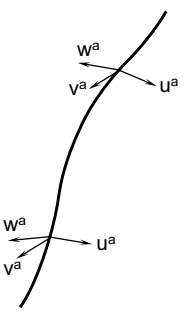
As a final example of these geometrical concepts, we return to the example of several observers passing between a pair of events p and q. In the figure, curve B has the smallest acceleration. Curve C appears to go away from B at large speed, accelerate quickly to turn around, and come back at a large speed. On the other hand, A never gets too far from B, but he is on a "bumpy ride", accelerating this way and that. Curve B is the longest. (Recall that a null curve has zero length. in A and C, the four-velocity tends to be "nearer the light cone", and hence these have shorter lengths. Thus, as a general rule, the observers who endure the most acceleration come back the youngest. A bumpy airplane flight ages you less than a smooth one. If you throw a rock into the air and catch it, the rock has aged more during this exercise than you have (for you were undergoing acceleration; the rock was not).

10. Rotation



There is more than one way for an observer to follow a given world-line. Specifically, an observer can rotate.

Rotation, like acceleration, is an unambiguous, absolute, local, measurable quantity. ("Rotation relative to what?", like "Acceleration relative to what?" is a fruitless question.) This remark is made explicit by the following measuring apparatus. Consider a solid sphere with many spikes sticking out of it. On each spike there is a small bead which is free to slide on the spike. The beads are connected to the sphere by springs so that, at equilibrium, the beads are all at some specified distance (e.g., three inches) from the sphere. Now, when such a sphere is rotated, the beads along the equator of the axis of rotation will move farther out from the sphere (centrifugal force). Thus, one can determine, at any instant, the state of rotation of the sphere by examining the configuration of the beads.



Choose three, orthogonally-directed, spikes on the sphere, and label them u, v, and w. Then, if this sphere is carried along a world-line, the way in which it is carried (regarding rotation) is described by specifying, at each point of the world-line, three vectors, u^a, v^a, w^a, which are orthogonal to each other and to the four-velocity.

In section 9, we related a physical quantity, acceleration, to the curvature of the world-line. We now wish to do a similar thing with rotation. The vectors u^a, v^a, and w^a determine the orientation of the sphere along its world-line. We wish to obtain an expression relating the rate of change of these vectors along the world-line to the rotation measured by the sphere. It suffices to ask the question: what equation must be satisfied by u^a, v^a, and w^a in order that the sphere record zero rotation? Then, the extent to which this equation is not satisfied (for the actual u, v, and w) is a measure of the rotation recorded by the sphere.

We now assert that *the sphere records zero rotation precisely when u^a, v^a, and w^a satisfy*

$$\begin{aligned} \xi^m \nabla_m u^a &= \xi^a (u_m A^m) \\ \xi^m \nabla_m v^a &= \xi^a (v_m A^m) \\ \xi^m \nabla_m w^a &= \xi^a (w_m A^m) \end{aligned} \tag{12}$$

where A^a is the acceleration (curvature) of the world-line. (Eqns. (12) define what is called *Fermi transport*.) We check that this assertion is reasonable. Note, firstly, that Eqns. (12) involve the first derivative of u, v, and w along the world-line. That is to say, they involve the "time rate of change of the orientation of the sphere", which, indeed, is a "rotation-like" quantity. Thus, in particular, if one specifies u^a, v^a, and w^a initially (i.e., at some point on the world-line), then Eqns. (12) determine these vectors uniquely along the entire world-line. (That's correct:

if you give me the orientation of the sphere at one time, and I then require that the sphere always record zero rotation, that determines its rotation thereafter.) Note, furthermore, that, if u, v, and w, are initially (i.e., at some point on the world-line) orthogonal, then according to Eqns. (12), they remain orthogonal:

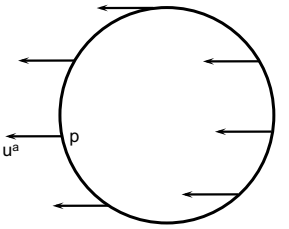
$$\begin{aligned}\xi^m \nabla_m (u_a v^a) &= u_a \xi^m \nabla_m v^a + v^a \xi^m \nabla_m u_a \\ &= u_a (\xi^a v_m A^m) + v^a (\xi_a u^m A_m) = 0.\end{aligned}$$

This, of course, is expected. Finally, note that a vector, transported according to Eqns. (12), remains orthogonal to the four-velocity:

$$\begin{aligned}\xi^m \nabla_m (u_a \xi^a) &= u_a \xi^m \nabla_m \xi^a + \xi^a (\xi^m \nabla_m u_a) \\ &= u_a (A^a) + \xi^a (\xi_a u_m A^m) = u_a A^a - u_m A^m = 0.\end{aligned}$$

Thus, Fermi transport has all the properties one would desire to describe the orientation of a sphere which records zero rotation.

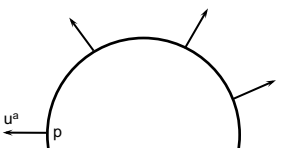
The following remarks are intended to give an intuitive picture of Fermi transport. Consider a circle in the plane. Let p be a point on the circle, and let u^a be a vector at p which points outward from the circle. Suppose we parallel transport this u around the circle. the resulting vector u^a is shown in the figure. Note that, while this u^a is initially orthogonal to the tangent vector to the circle, this property is not preserved under parallel transport. If, on the other hand, v^a were a vector at p which sticks out of the paper, the parallel transport of v^a around the circle would keep v^a orthogonal to the tangent vector to the circle. Why this difference between between u^a and v^a? Note that the curvature of the circle, C^a, is a vector pointing inward toward the center. Thus, since the circle "curves in (minus) the direction of u^a", u^a does not remain orthogonal to the tangent to the circle. Since the curvature of the circle is orthogonal to v^a, v^a does remain orthogonal to the tangent vector.



Intuitively, in Fermi transport one "parallel transports the vector a little way along the curve, then projects orthogonal to the tangent vector, parallel transports a little further, then projects again, etc." Under "parallel transport for a little way" the vector "loses its orthogonality with the tangent vector" by an amount determined by $u^a A_a$ (as we saw above; $u^a A_a$ is nonzero. and u^a loses orthogonality under parallel transport, $v^a A_a$ is zero, and v^a does not lose orthogonality under parallel transport). Thus, to "restore orthogonality" one must correct parallel transport by an amount involving $u^a A_a$. Thus, the term on the right in (12). The ξ^a appears on the right because we want to "correct parallel transport of u^a by a multiple of the tangent vector". In other words, we want to project orthogonally to ξ^a (an operation which is accomplished by adding an appropriate multiple of the tangent vector). In short, Fermi transport is parallel transport with a correction term to keep the transported vector orthogonal to the four-velocity.

Note, in particular, that *for a geodesic* (i.e., for $A^a = 0$) *Fermi transport and parallel transport coincide.*

Finally, the figure on the right represents the result of Fermi transport of u^a (given at p) around the circle. That

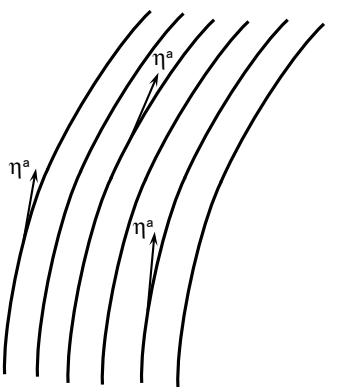


this is the right answer should be clear from the remark above. For v^a (the vector which sticks out of the paper), Fermi transport and parallel transport coincide (for $v^a A_a = 0$. See (12).) (In general, parallel transport about a closed curve leaves a vector unchanged only when the space is flat. Fermi transport about a closed curve does not leave a vector unchanged - even in flat space.)

11. Dust

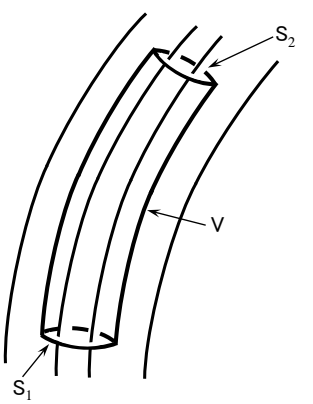
The source of a gravitational field is matter. In Newtonian theory, this fact is expressed by the equation $\nabla^2\varphi = 4\pi G\rho$. Thus, all the various types of matter have in common, in Newtonian theory, the feature that they determine a ρ, the matter density, which generates gravitation. In general relativity, matter generates gravitational field via a certain tensor field T_{ab}, called the *stress-energy tensor* of the matter. (So, T_{ab} is analogous to the Newtonian ρ.) We now begin our study of various types of matter to see how each, ultimately, defines a stress-energy tensor.

Consider a cloud of dust, i.e., an assembly of a large number of very small, very light particles which do not directly interact with each other. In terms of space-time, we would represent this dust cloud by drawing the world-lines of all the dust particles. This congruence of timelike curves would fill a certain region of space-time.



The idea is to describe this situation in terms of certain tensor fields in space-time. At each event of space-time inside the dust cloud, let η^a be the four-velocity of the dust particle there. In this way, we obtain a unit ($\eta^a\eta_a = -1$) vector field η^a in the region of space-time occupied by the dust. This η^a is called the (four-) *velocity field* of the dust. This is the first of two tensor fields describing the dust.

The second is the *density*, ρ, of the dust. At a point p of the dust cloud, draw a small 3-surface S orthogonal to η^a. Denote by V the volume (in sec^3) of this surface. Then ρ is defined as Nm/V, where N is the number of dust particle world-lines which pass through S, and m is the mass of each dust particle. (So, ρ has units grams/sec^3.) Note that this ρ is just the usual "mass density" as would be determined by an observer traveling along a world-line of the dust. This ρ is a scalar field in the region occupied by the dust. (More precisely, we are thinking of a limit in which the number of dust particles increases, and the mass of each one approaches zero, such that ρ remains finite in the limit.)



Note: density of water $= 3 \times 10^{31}$ g/sec^3.

We assume that ρ and η^a are smooth fields. (Rule of thumb: Whenever smoothness is meaningful for a physical quantity, assume it.)

What equations would we expect the fields ρ and η^a to satisfy? The first expresses the fact that the dust particles are non interacting, i.e., that they are free particles. The corresponding equation, clearly, is $\eta^m\nabla_m\eta^a = 0$. That is, the dust particles travel along geodesics. The second equation expresses the conservation of energy (or, since all particles have a fixed mass m, of dust particles). Consider a tubular volume V whose sides are parallel to η^a and whose faces are orthogonal

to η^a. The total amount of mass (in dust) entering this tune is

$$\int_{S_1} (\rho\eta_a)\mathrm{d}S^a$$

where S_1 is the bottom face of the tube, and $\mathrm{d}S^a$ is the volume element on this face (directed outward). Similarly, the total mass in dust leaving the tube is

$$-\int_{S_2} (\rho\eta_a)\mathrm{d}S^a$$

Conservation of dust particles requires

$$\left(\int_{S_1} + \int_{S_2}\right) (\rho\eta_a)\mathrm{d}S^a = 0$$

or, what is the same thing,

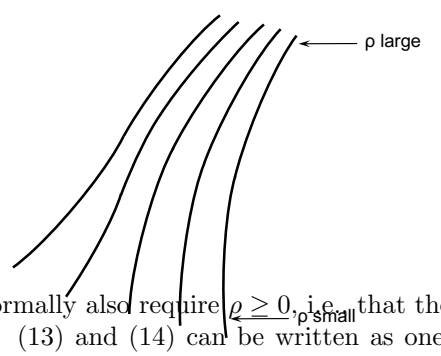
$$\int_{\partial V} (\rho\eta_a)\mathrm{d}S^a = 0$$

In the second surface integral, we have included the sides of the tube, but these make no contribution, since $\eta_a dS^a = 0$ there. By Gauss' law, we now have

$$\int_V \nabla_a(\rho\eta^a)\mathrm{d}V = 0$$

Finally, since the volume V is arbitrary, we have $\nabla_a(\rho\eta^a) = 0$. Intuitively, this equation states that, when the world-line of the dust-particles draw closer together, the density ρ increases by a corresponding amount. (Write the equation in the form $\eta^a\nabla_a\rho = -\rho\nabla_a\eta^a$ to see this.) This, of course, in what we would expect to happen from conservation of dust particles.

All the discussion above can be regarded as merely motivation. When we say "dust", we mean merely a pair of fields, ρ and η^a, satisfying



$$\eta^m\nabla_m\eta^a = 0 \qquad (13)$$

$$\nabla_a(\rho\eta^a) = 0 \qquad (14)$$

(and, of course, $\eta^a\eta_a = -1$. One would normally also require $\rho \geq 0$, i.e. that the dust particles have positive mass.) Eqns. (13) and (14) can be written as one. Multiplying (13) by ρ, and (14) by η^a, and adding:

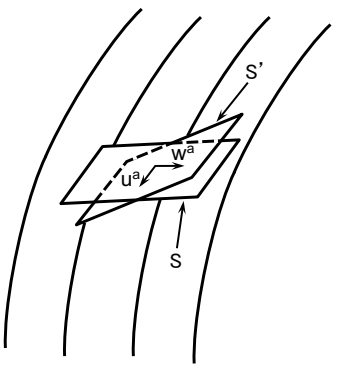
$$\rho\,\eta^m\nabla_m\eta^a + \eta^a\nabla_m(\rho\,\eta^m) = 0 \qquad (15)$$

Conversely, contracting (15) with η_a (using $\eta_a A^a = 0$), we obtain (14). Then (15) gives (13). Thus, Eqn.(15) is completely equivalent to (13) and (14). Finally, note that (15) can be written

$$\nabla_b T^{ab} = 0 \qquad (16)$$

where we have set $T^{ab} = \rho\,\eta^a\eta^b$. This T^{ab} is called the *stress-energy* of the dust. The behavior of the dust is then completely described by (16), which is called the *conservation equation*.

We next give direct physical interpretation of T^{ab}. Let an observer pass through our dust cloud, and let his four-velocity be ξ^a. We claim that the quantity $-\xi_b T^{ab}$ is the four-momentum density of the dust, as seen by this observer. (So, in particular, $\xi_a \xi_b T^{ab}$ is the observed energy density.) Suppose first that $\xi^a = \eta^a$. Then $-\eta_b T^{ab} = \rho\, \eta^a$. Evidently (from the definition of ρ), this is the four-momentum density of the dust as seen by an observer moving with the dust. (Recall that $m\eta^a$ is the four-momentum of a single dust particle. So, $\rho\eta^a = (N/V)m\eta^a$ is the four-momentum density as seen by this observer.) Next, suppose general ξ^a Then $-\xi_b T^{ab} = (-\xi_b\eta^b)\rho\eta^a$. The second factor on the right is (as noted above) the four-momentum density seen by an observer following the dust. We shall show below that $(-\xi_b\eta^b)^{-1}$ is precisely the ratio of the density of dust particles seen by the ξ-observer to the density seen by the η-observer. This having been done, the interpretation claimed above for $-\xi_b T^{ab}$ will be clear.



Fix an event p on the world-line of our observer. Let S be a small cubical spacelike 3-surface at p which is orthogonal to η^a. By "tipping" the boundary of S, construct a second cubical spacelike 3-surface S' at p, this one orthogonal to ξ^a. Clearly, the same number of dust trajectories pass through S and S'. Thus, we have only to show that $V'/V = (-\xi_a\eta^a)^{-1}$, where V' and V are the areas (in sec^3) of S' and S, respectively. Let the sides of the cube S be described by unit vectors u^a, v^a and w^a at p, where these vectors are orthogonal to each other and to η^a. Furthermore, let w^a lie in the (2-dimensional) plane of ξ^a and η^a. Then the sides of S' are described by the vectors u^a, v^a, and $w^a - \eta^a(\xi^m w_m)(\xi^n\eta_n)$ i.e., we add the appropriate multiple of η^a to w^a so the result will be orthogonal to ξ^a). But, since w^a is unit, orthogonal to η^a and a linear combination of ξ^a and η^a, we have $w^a = ((\xi^m\eta_m)^2 - 1)^{-1/2}(\xi^a + (\xi^m\eta_m)\eta^a)$. Hence, $\xi^m w_m = ((\xi^m\eta_m)^2 - 1)^{1/2}$. We conclude that the edges, of S' are described by u^a, v^a, and $w^a - \eta^a(\xi^a\eta_m)^{-1}((\xi^n\eta_n)^2 - 1)^{1/2} = w'^a$. We now have three vectors describing the edges of S'. Clearly, $V'/V = (w'^a w'_a)^{1/2} = (-\xi^n\eta_n)^{-1}$. This completes the proof. (This argument is well-known in special relativity. it is called the "transformation law for a density under a boost", and the formula is explained by saying that "one of the three spatial dimensions suffers a Lorentz contraction".)

It is true quite generally (as shown explicitly above) that $-\xi_b T^{ab}$ *can be interpreted as the four-momentum density of the matter described by* T^{ab}*, according to an observer with four-velocity* ξ^b. Thus, we can think of T^{ab} as having two indices because "one will be a four-momentum index, while the other must be contracted with ξ_b in order that the correct density transformation between observers obtains".

In astrophysics, the stars in a galaxy, or the galaxies in the universe, are often treated as "dust particles". Dust is one of the simplest sources.

12. Perfect fluids

The second type of matter we shall discuss are the perfect fluids (e.g., water, air). Dust, as we shall see, may be regarded as a special case of a perfect fluid.

Consider a fluid residing in our space-time. As usual, we want a description in terms of tensor fields. Imagine tagging in some way a small volume of fluid. In space-time, this volume would have some world-lines. Let η^a be the unit tangent vector to this world-line. Repeating, at each event of space-time in the fluid, we obtain a unit timelike vector field η^a. This is called the (four-) *velocity field* of the fluid. At each event in the fluid, an observer passing though that event with four-velocity η^a could measure the density (g/sec^3) of the fluid. Thus, we have the *density* ρ of the fluid a scalar field in space-time. Finally, an observer with four-velocity η^a could determine a quantity p, the *pressure* of the fluid. Of course, p is a scalar field in space-time. (The units of pressure are g/sec^3. Thus, 1 dyne /cm^2 $= 3 \times 10^{10}$ g/sec^3.)

Thus, a perfect fluid is defined by three fields, η^a, ρ, and p, in space-time. One would normally require that ρ be non-negative. (On the other hand, many fluids, e.g., water, can exert negative pressure.) The specification of what fluid is under consideration is carried out by giving the pressure p as a function of ρ, $p(\rho)$. (For example, if you say what you want the density of air to be, I can tell you what pressure must be exerted on a sample of air to achieve that density. The function $p(\rho)$ would, of course, be quite different for water.) This $p(\rho)$ is called the *equation of state* of the fluid. (Of course, one could consider more complicated fluids. One could, for example, introduce temperature, entropy density, etc., and write more complicated equations of state. While these more complicated fluids are important for certain applications, their introduction is straightforward, and they add nothing fundamental to the study of general relativity. We shall not consider then.)

We could now proceed with fluids as we did before with dust. We could agree on some reasonable equations to be satisfied by η^a, ρ , and p. Instead, we shall simply write down the stress-energy tensor for a perfect fluid, and then verify that its conservation leads to reasonable physics.

The stress-energy tensor for a perfect fluid is given by

$$T^{ab} = (\rho + p)\ \eta^a \eta^b + p\ g^{ab}.$$

It follows immediately that dust is a special case of a perfect fluid – the case with equation of state $p(\rho) = 0$. The conservation of stress-energy, $\nabla_b T^{ab} = 0$, gives for a perfect fluid:

$$(\rho + p)\eta^b \nabla_b \eta^a + (g^{ab} + \eta^a \eta^b)\nabla_b p + \eta^a p \nabla_b \eta^b + \eta^a \nabla_b (\rho \eta^b) = 0 \qquad (17)$$

In order to extract some physics from (17), we project parallel and orthogonal to η^a. (This procedure is useful, as a general rule, when dealing with fluids.) Contracting (17) with η^a, we have

$$\nabla_b (\rho \eta^b) = -p \nabla_b \eta^b. \qquad (18)$$

For dust, the right side of (18) vanishes. In this case, our interpretation of (18) was that it gives conservation of mass (for the dust). Thus, the presence of a non vanishing right side of (18) indicates that mass is no longer conserved. How shall we interpret this physically? Consider (in Newtonian physics) a sample of gas of volume V, and at pressure p. Suppose this gas is allowed to expand slightly, so its

volume becomes $V + \mathrm{d}V$. Then, in this process, an energy $p\mathrm{d}V$ is extracted from the gas. (Let the gas be in a cylinder with a piston of area A. Then $\mathrm{d}V = A\mathrm{d}l$, where $\mathrm{d}l$ is the distance the piston moves. The pressure force on the piston is pA. Hence, the energy extracted is force $\times$ distance $pA \times \mathrm{d}l = p\mathrm{d}V$.) Recall that $\nabla_b \eta^b$ represents the rate at which a small volume of fluid is expanding. The interpretation of (18) is now clear: it states that the energy pumped into the fluid (by pressure acting against volume changes) goes into increasing the mass density of the fluid. (If you compress a sample of gas, it gets heavier because of the energy you are putting into it. $E = mc^2$, and all that.) Thus, Eqn. (18) is conservation of mass-energy.

The other half of Eqn. (17) is obtained by projecting orthogonal to η_a. This is accomplished by contracting it with one index of the projection operator, $h_{ac} = g_{ac} + \eta_a \eta_c$, orthogonal to η^a. Thus, we obtain

$$(\rho + p)\ \eta^b \nabla_b \eta^a = -(g^{ab} + \eta^a \eta^b) \nabla_b p. \tag{19}$$

Again, in the case of dust, the right side of (19) vanishes, and the equation then represents the statement that the dust particles are free particles. Hence, in the present case, we wish to interpret the right side of (19) as a force on small volume elements of the fluid. Note that the right side of (19) can be written $h^{am} \nabla_m p$, where h^{am} is the projection operator orthogonal to the fluid. In other words, the right side of (19) represents a pressure force on fluid elements. Recall that, in Newtonian mechanics, the gradient of the pressure gives a difference in the force on opposite faces of a small fluid element, and hence a net force on that element. The situation is the same here, except that $\nabla_b p$ is the gradient in space-time, and this gradient must be projected orthogonal to η^a to obtain the force. Note also that the coefficient of the acceleration on the left in (19) is $(\rho + p)$. This quantity thus represents the "effective inertial mass density" of the fluid.

We have in (19) an example of a force in general relativity. One begins with a vector field (in the example above, $\nabla_b p$). Then, the force on a particle is obtained by projecting that vector orthogonal to the four-velocity of the particle.

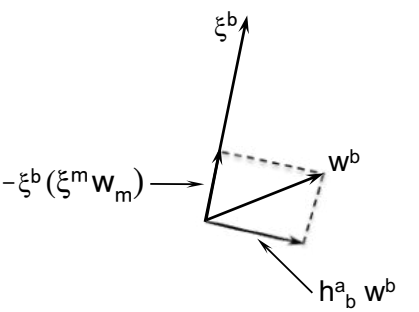
To summarize, a perfect fluid is characterized by fields, η^a, ρ, and p, which define a stress-energy T^{ab}. The component of $\nabla_b T^{ab} = 0$ parallel to η^a gives conservation of mass-energy, the component orthogonal to η^a conservation of momentum.

13. Electromagnetic Fields: Decomposition by an Observer

We now begin our treatment of the third important type of "matter": electromagnetic fields.

An *electromagnetic field* is a (smooth) antisymmetric tensor field, F_{ab} (= $F_{[ab]}$), on space-time M. An individual observer resolves this single object – the electromagnetic field – into the separate electric and magnetic fields seen by him. Our first task is to see how this resolution comes about. Let our observer have four-velocity ξ^a. Recall that $h_{ab} = g_{ab} + \xi_a \xi_b$ is the projection operator orthogonal to ξ^a. That is, if w^b is w is any vector (at a point of the observer's world-line), then$h^a{}_b w^v$ is the component of w^b orthogonal to ξ^a. (Proof: For any w^b we have $w^a = h^a{}_b w^b - \xi^a(\xi^m w_m)$. But the second term on the right is the component of w^b parallel to ξ^a.) Thus, in particular we have $h^a{}_m h^m{}_b = h^a{}_b$ (the component orthogonal to ξ of the component orthogonal to ξ of a vector is the component orthogonal to ξ of that vector), and $\xi^b h^a{}_b = 0$ (the component of ξ^b orthogonal to ξ is zero).

To resolve F_{ab} into "spatial tensors" for our observer, we project the indices of F_{ab} parallel and orthogonal to ξ^a. Thus, we obtain four tensors: $F_{mn}\xi^m\xi^n$, $F_{mn}\xi^m h^n{}_a$, $F_{mn}h^m{}_a\xi^n$, and $F_{mn}h^m{}_a h^n{}_b$. Since F_{ab} in antisymmetric, the first vanishes, and the second equals minus the third. The third is called the *electric field*:



$$E^a = F_{ab}\,\xi^b. \tag{20}$$

(Note that $F_{am}\,\xi^m = F_{mn}\xi^m h^n{}_a$) Note that the determination of an electric field from the electromagnetic field involves a choice of an observer (i.e., of his four-velocity). The remaining piece of F_{ab} is $F_{mn}\ h^m{}_a\ h^n{}_b$, a spatial, antisymmetric tensor. To express this piece more simply, let ϵ_{abcd} be the alternating tensor on M (i.e., the (unique up to sign) tensor field defined by $\epsilon_{abcd} = \epsilon_{[abcd]}$ and $\epsilon_{abcd}\epsilon^{abcd} = -24$). Then $\epsilon_{abc} = \epsilon_{abcd}\,\xi^d$ is the spatial alternating tensor as seen by our observer (i,e., it satisfies $\epsilon_{abc} = \epsilon_{[abc]}$, $\epsilon_{abc}\epsilon^{abc} = 6$, and $\xi^a\ \epsilon_{abc} = 0$).

A spatial anti symmetric W_{ab} for our observer (i.e,. satisfying $W_{ab} = W_{[ab]}$, and $\xi^a\ W_{ab} = 0$) is then completely described by the corresponding vector, $\epsilon^{abc}\ W_{bc} = W^a$. Note that W^a is also a spatial vector (for our observer), and that W_{ab} can be recovered from W^c: $W_{ab} = \frac{1}{2}\epsilon_{abc}\ W^c$. The spatial vector which gives $F_{mn}\ h^m{}_a\ h^n{}_b$ is thus

$$B_a = \frac{1}{2}\ \epsilon_{abcd}\ \xi^b\ F^{cd}. \tag{21}$$

This B_a is called the *magnetic field*.

To summarize, the whole story is described by a single tensor field, the electromagnetic field. An observer resolves this F_{ab} using his four-velocity, into a pair of vectors, E_a and B_a, orthogonal to his four-velocity. The E_a and B_a depend on the observer. (In terms of components, the six independent components of F_{ab} become three in E_a and three in B_a.)

We now reverse the procedure of Eqns. (20) and (21): we express F_{ab} in terms of E_a and B_a. The equation is

$$F_{ab} = 2\xi_{[a}E_{b]} - \epsilon_{abcd}\ \xi^c\ B^d. \tag{22}$$

To verify this one substitutes (22) into (20), and (22) into (21), to obtain identities.

An electromagnetic field produces a force on a charged particle. Consider a particle with charge e, and let η^a be the four-velocity of the particle. Then the force on this particle is the right side of:

$$m\ \eta^b\ \nabla_b\ \eta^a = e\ F^{ab}\ \eta_b. \tag{23}$$

Thus, Eqn. (23) is the equation of the motion of the particle (m is its mass). Note that the right side is orthogonal to η_a (as we would demand, since the left side automatically is), since F_{ab} is skew. Eqn. (23) gives immediately the equation of motion for charged dust. Suppose each particle of the dust has charge e, and let μ be the charge density of the dust (defined exactly as the mass density. Thus, in particular $\rho/\mu = m/e$.) Then, in this case, Eqn. (13) would be replaced by

$$\rho\ \eta^b\ \nabla_b\ \eta^a = \mu\ F^{ab}\ \eta_b. \tag{24}$$

Our next objective is to recover from (23), the usual classical formula for the force on a charged particle. Through a point of space-time, let there pass an observer with four-velocity ξ^a, and a charged particle with four-velocity η^a. We suppose (to simplify the discussion) that η^a differs infinitesimally from ξ^a, i.e., we set $\eta^a = \xi^a + v^a$, where v^a is small. (It follows from the fact that η^a and ξ^a are unit vectors that $\xi^a\ v_a = 0$.) This v^a is the velocity of the charged particle, as seen by the observer. Now, our observer decomposes F_{ab} into E_a and B_a as described above. Hence, he can express the force on the particle in terms of E_a, B_a, and v_a. Unfortunately, this force, $e\ F_{ab}\ \eta^b$, is orthogonal to η^a (the four-velocity of the particle),and not to ξ^a (the observer's four-velocity). However, we can project it orthogonal to the observer's four-velocity using $h_{ab} = g_{ab} + \xi_a\xi_b$. Thus, the force is given by (first order in v^a)

$$\begin{aligned} e\ h^m_a\ \eta^n F_{mn} &= e(\delta^m_a + \xi_a\ \xi^m)(\xi^n + v^n) \\ &\quad (2\ E_{[m}\ \xi_{n]} - \epsilon_{mncd}\ \xi^c\ B^d) \\ &= e\ E_a + e\ \epsilon_{abc}\ v^b\ B^c \qquad (\text{where}\ \ \epsilon_{abc} = \epsilon_{abcd}\ \xi^d) \end{aligned}$$

In standard three-dimensional vector notation, this equation would be written $\vec{F} = e\ \vec{E} + e\ \vec{v} \times \vec{B}$, the usual force equation.

In everyday life in general (or special) relativity, one never uses the decompositions of things as seen by an observer. These are given here merely for motivation – as an opportunity to recover some familiar formulae. Thus, eliminating all this motivation, there are only two statements in this entire section. The electromagnetic field is a skew tensor field, F_{ab}, on space-time. The force on a particle of charge e is given by the right side of Eqn. (23).

14. Maxwell's equation

We have now introduced electromagnetic fields, and the equation (23) for the action of this field on charged particles. the next step is to describe the effects of charges on the fields. These, Maxwell's equations, express the production of the electromagnetic fields by charges and currents.

Maxwell's equations are

$$\nabla^{[a} F^{bc]} = 0. \tag{25}$$

$$\nabla_b F^{ab} = J^a. \tag{26}$$

Here, F_{ab} is the electromagnetic field, and J_a is a certain vector field in space-time called the *charge-current density*. This charge current density is obtained from the field describing the charged matter. We begin with the physical interpretation of J_a. Let an observer have four-velocity ξ^a. Then $(-J_a\ \xi^a)$ is the *charge-density* as seen by our observer, while $h_a{}^m\ J_m$ is the *current-density* as seen by this observer. For example, for charged dust, the charge-current density is given by

$$J^a = \mu\ \eta^a,$$

where μ is the charge-density of the dust. For example, for an observer with four-velocity η^a – i.e., an observer traveling with the dust – the charge-density he sees is $(-J_a\ \eta^a) = -\mu\ \eta_a \eta^a = \mu$. He sees zero current-density (as one would expect, he moves with the dust. On the other hand, an observer with four-velocity different from that of the dust sees a current-density (physically, he sees all those charged particles rushing by).

An important equation on J_a follows immediately from (26). Taking the divergence of this equation, we have

$$\nabla_a\ J^a = \nabla_a \nabla_b\ F^{ab}.$$

But, from the definition of the Riemann tensor (commuting derivatives),

$$\nabla_a \nabla_b\ F^{cd} = -\nabla_b \nabla_a\ F^{cd} = R_{ab}{}^c{}_m\ F^{md} + R_{ab}{}^d{}_m\ F^{cm}. \tag{27}$$

Contracting (27) over "a" and "c", and over "b" and "d", using the fact that F_{ab} is antisymmetric, we see that the right side of the preceding equation vanishes. That is, $\nabla_a J^a = 0$. What this means should be clear from Sect. 11. It is conservation of charge. Thus, the source for the electromagnetic field, the charge-currant density, is divergence-free, expressing conservation of charge.

Finally, we verify that (25) and (26), when expressed in terms of E_a and B_a (as determined by some observer), become the usual formulae for Maxwell's equations. Unfortunately, Eqns. (25) and (26) involve derivatives; hence, they are difficult to express at a point in terms of the observations of a single observer. We therefore introduce a field of observers.

Let our space-time be that of special relativity (i.e., let the Riemann tensor vanish), and let ξ^a be a constant ($\nabla_a\ \xi^b = 0$, unit, timelike vector field in space-time. (There exists no such vector field in general in general relativity. Proof: $0 = \nabla_{[a}(\nabla_{b]}\ \xi^c) = \nabla_{[a}\nabla_{b]}\ \xi^c = \frac{1}{2} R_{ab}{}^c{}_d\ \xi^d$.) We can now carry out the decomposition of F_{ab} into E_a and B_a at each point of space-time. We also decompose our derivative operator: let a dot denote $\xi^a\ \nabla_a$ (the time derivative), and let $D_a = h_a{}^b \nabla_b$ (the spatial derivative). Thus, we now have a special symbol for the spatial and

temporal parts (with respect to our field of observers, i.e., with respect to ξ^a) of every quantity appearing in (25) and (26).

Contacting Eqns. (26) with ξ^a, we have $\xi_a \nabla_b F^{ab} = \xi_a J^a$. But $\xi_a \nabla_b F^{ab} = \nabla_b(\xi_a F^{ab}) = -\nabla_b E^b = -h^{ab}\nabla_a E_b = -D_b E^b$. Furthermore, $\xi_a J^a = -\kappa$, where κ is the charge-density (as seen by our field of observers). Hence, Eqn. (26), contracted with ξ_a, yields

$$D_a E^a = \kappa. \tag{28}$$

This is the first (of four) Maxwell equations. The second Maxwell equation is obtained by projecting the index of Eqn. (26) orthogonally to ξ^a: $h_{ac}\ \nabla_b F^{cb} = h_{ac} J^c$. But $h_{ac} J^c = j_a$ the current-density (as seen by our field of observers). Furthermore,

$$\begin{aligned} h_{ac}\nabla_b F^{cb} &= \nabla_b(h_{ac}F^{cb}) = \nabla_b(-E_a\xi^b - \epsilon_a{}^{bcd}\xi_c B_d) \\ &= -\xi^b\nabla_b E_a - \epsilon_{abcd}\xi^c\nabla^b B^d = -\xi^b\nabla_b E_a + \epsilon_{abc}D^b B^c. \end{aligned}$$

Hence, projecting the index of (26) orthogonally to ξ^a, we obtain

$$\dot{E}_a - \epsilon_{abc}D^b B^c = -j_a. \tag{29}$$

In vector notation, the second term on the left would be written $-\vec{\nabla} \times \vec{B}$. Eqn. (29) is the second Maxwell equation.

The final two Maxwell equations come from (25). The calculation is almost identical to that above, so we do not give it in detail. Write Eqn. (25) in the form $\epsilon_{abcd}\nabla^b F^{cd} = 0$. Contracting with ξ_a, and projecting the index "a" orthogonally to ξ_a, we obtain, respectively,

$$D_a B^a = 0. \tag{30}$$

$$\dot{B}_a + \epsilon_{abc}D^b E^c = 0. \tag{31}$$

Eqns. (28), (29), (30), and (31) will be recognized as Maxwell's equations written out in "space and time" form. We have shown that this usual form for the equations follows from the much simpler ("space-time") equations (25) and (26).

Once again, it is true that almost everything in this section (in particular, everything involving space and time decompositions relative to an observer) can be regarded as merely motivation for what is a very simple situation in terms of space-time. Eliminating all motivation, what remains is the following. There is a vector field J_a in space-time, the charge current-density. This vector field acts as a source for electromagnetic field, in the sense that F_{ab} is required to satisfy Maxwell's equation, (25) and (26). It follows, in particular, from (26) that the divergence of J^a vanishes. (We shall never again use Eqns. (28), (29), (30), and (31).)

15. Stress Energy of the Electromagnetic Field

We have now written the complete set of equations describing the electromagnetic field. These give the effect of an electromagnetic field on charged matter (Eqn. (23)), and the effect of charged matter on the electromagnetic field (Eqns. (25) and (26)). We have seen in the cases of dust and fluids that the fields which describe the matter can be combined to give a stress-energy tensor, T^{ab} , which is conserved, and which is such that $\xi_a\xi_b T^{ab}$ is the energy density of the matter as seen by an observer with four-velocity ξ^a. We now show that an electromagnetic field defines a stress-energy with similar properties.

The stress-energy tensor of the electromagnetic field is given by the following equation:

$$T^{ab} = F^{am}F^{b}{}_{m} - \frac{1}{4}g^{ab}F^{mn}F_{mn}. \tag{32}$$

As usual, we begin by seeing what (32) means for our favorite observer – the one with four-velocity ξ^a. First note that, substituting (22), we have

$$\begin{aligned} F^{mn}F_{mn} &= (2\xi^{[m}E^{n]} - \epsilon^{mncd}\xi_c B_d)(2\xi_m E_n - \epsilon_{mnrs}\xi^r B^s) \\ &= 2(\xi^m\xi_m)(E^n E_n) - 2(\xi^c\xi_c)B^d B_d \\ &= 2(B^c B_c - E^c E_c), \end{aligned}$$

where we have used $\epsilon_{abmn}\epsilon^{cdmn} = -4\delta_a{}^{[c}\delta_b{}^{d]}$. Hence, from (32),

$$\begin{aligned} \xi_a\xi_b T^{ab} &= (\xi_a F^{am})(\xi_b F^{b}{}_{m} - \frac{1}{4}\xi^a\xi_a F^{mn}F_{mn} \\ &= E^m E_m + \frac{1}{2}(B^m B_m - E^m E_m) \\ &= \frac{1}{2}(E^m E_m + B^m B_m) \end{aligned} \tag{33}$$

Similarly,

$$\begin{aligned} -h_{ac}\xi_d T^{cd} &= -h_{ac}\xi_d[F^{cm}F^{d}{}_{m} - \frac{1}{4}g^{cd}F^{mn}F_{mn}] \\ &= -h_{ac}\xi_d F^{cm}F^{d}{}_{m} = h_{ac}E_m F^{cm} \\ &= h_{ac}E_m[2\xi^{[c}E^{m]} - \epsilon^{cmrs}\xi_r B_s] \\ &= -h_{ac}E_m\epsilon^{cmrs}\xi_r B_s = \epsilon_{abcd}E^b B^c \xi^d \\ &= \epsilon_{abc}E^b B^c. \end{aligned} \tag{34}$$

The right sides of (33) and (34) will be recognized from classical electromagnetism. The right side of (33) is what is called the energy density of the electromagnetic field. The right side of (34), written $\vec{E}\times\vec{B}$ in vector notation, is called the Poynting vector, and is interpreted as the momentum density of the electromagnetic field.

Thus, if we take for (32) the stress-energy of the electromagnetic field, we obtain the usual conclusions: $-T^{ab}\xi_b$ is the four-momentum density of the electromagnetic field, as seen by an observer with four-velocity ξ^b.

Note that T^{ab} in (32) is symmetric and trace-free (i.e., $T^{ab} = T^{(ab)}$, and $T^m{}_m = 0$).

As a consequence of the equations satisfied by the fields describing dust, the stress-energy of dust is conserved. As a consequence of the equations satisfied by the fields describing a perfect fluid, the stress-energy of the fluid is conserved. Is the stress-energy of the electromagnetic field, (32), conserved? We find out:

$$\begin{aligned} \nabla_b T^{ab} &= F^{am}\nabla_b F^b{}_m + F^b{}_m \nabla_b F^{am} - \frac{1}{2}F_{mn}\nabla^a F^{mn} \\ &= F^{am}\nabla_b F^b{}_m - \frac{3}{2}F_{mn}\nabla^{[a}F^{mn]} \\ &= -F^{am}J_m. \end{aligned} \tag{35}$$

Where, in the last step, we have used Maxwell's equations, (25) and (26).

Consider first the case when the charge-current J^a vanishes. Then, from (35), we indeed have $\nabla_b T^{ab} = 0$. But, when J^a is nonzero, we do not in general have conservation of the electromagnetic stress-energy. What is going on? Recall that the conservation of stress-energy represents, physically, the local conservation of energy and momentum. An electromagnetic field without sources has an energy and momentum (according to an observer) which is conserved. But, when we have charged particles around, the electromagnetic field has the possibility of transferring energy and momentum to these particles. Hence, it should not be the stress-energy of the electromagnetic field alone which is conserved. Instead, it should be the total stress-energy – the sum of that due to the fields and that due to the sources – which is conserved.

We illustrate the remarks above for the case of charged dust and electromagnetic fields. The fields are the four-velocity of the dust, η^a, the mass density of the dust, ρ, the charge density of the dust, μ, and the electromagnetic field, F_{ab}. These fields satisfy conservation of mass for the dust ($\nabla_a(\rho\eta^a) = 0$), equation of motion for dust particles (Eqn. (24)), and Maxwell's equations (Eqns. (25) and (26)). For the stress-energy of this system, we take the sum of the dust and electromagnetic contributions:

$$T^{ab} = \rho\eta^a\eta^b + F^{am}F^b{}_m - \frac{1}{4}g^{ab}F^{mn}F_{mn}.$$

Hence,

$$\nabla_b T^{ab} = \eta^a\nabla_b(\rho\eta^b) + \rho\eta^b\nabla_b\eta^a + F^{am}\nabla_b F^b{}_m - \frac{3}{2}F_{mn}\nabla^{[a}F^{mn]}.$$

The first term on the right vanishes by conservation of mass for the dust. The second term cancels the third by (24) and (26). The fourth term vanishes by (25). Hence, we obtain $\nabla_b T^{ab} = 0$, as expected.

Once again, this entire section can be regarded as consisting of motivation, except for one point. The stress-energy of the electromagnetic field is given by Eqn. (32).

16. Stress-Energy: A Summary

There are other things in the world besides dust, fluids, and electromagnetic fields. For example, there are solids (matter with internal stresses), Dirac fields (which describe electrons), etc. In the preceding five sections, we have treated just three cases – dust, fluids and electromagnetism – in detail. (These three, at least in part because of their simplicity, are the most common.) Certain broad features of the discussion of the last five sections are, it is assumed, valid for all matter and fields which occur in Nature. For all matter and all fields which have been so far investigated, these broad features have been found to hold. As a summary, we now list these features.

Each type of matter is described by a certain collection of smooth tensor fields in space-time. These fields satisfy certain differential equations. The equations for the fields of one type of matter may involve the fields of another type of matter. In this case, the two types of matter are considered as interacting with one another. For each type of matter, there is an expression for a tensor field T^{ab} in terms of the fields of that matter. This T^{ab}, the *stress-energy* (of that type of matter), is symmetric. Now suppose we assemble a list of various type of matter, where this list has the following property: the equations for all the fields for all the types of matter in this list involve only the fields for the type of matter in the list. (Physically, suppose that each type of matter in the list interacts only with types of matter which appear in the list.) Then T^{ab}, the sum of the stress-energies of all the types of matter in the list, is the total stress-energy of this system of interacting types of matter. Then $\nabla_b T^{ab} = 0$, i.e., the total stress-energy is conserved.

Let an observer have four-velocity ξ^a. Let T^{ab} be the stress-energy of one particular type of matter. Our observer interprets $-\xi_b T^{ab}$ as the four-momentum density of this type of matter, so, in particular, he interprets $\xi_a \xi_b T^{ab}$ as the mass (energy) density of that type of matter. This observer regards conservation of the total stress-energy (of some interacting types of matter) as representing local conservation of energy and momentum. That the stress-energy of each individual type of matter is not conserved is no surprise to him. He says that the various types of matter exchange energy and momentum (from one type to another) through their interaction.

17. The Optical Limit

It is known that light and electromagnetic fields are by no means disjoint phenomena in physics. In fact, it is known that what we call light is merely waves in the electromagnetic fields. Thus, in particular, all the properties of light follow already from Maxwell's equations ((25) and (26)). Associated with light are diffraction phenomena – phenomena which become important when one looks at distance comparable with the wavelength of the light. In the limit of very small wavelength (i.e., when the wavelength is much smaller than all other relevant distances), the behavior of light is that described by geometrical optics (i.e., light goes in straight lines, etc.). We shall call this (short wavelength) limit in the *optical limit*.

In Sect. 3, we used light to obtain the light-cone structure of space-time. By "light" in that section, we meant light (i.e., electromagnetic waves) in this optical limit. From the way we defined things in Sect. 3, it follows that light rays travel on null curves. But now we have Maxwell's equation. We now ask: is the behavior of electromagnetic waves (as determined by Maxwell's equations) such that, in the optical limit, these waves (light rays) do indeed travel on null curves? We shall here not only answer this question affirmatively, but, furthermore, obtain additional properties of light rays. We shall here take the optical limit of Maxwell's equations.

We consider an electromagnetic field of the form

$$F_{ab} = \overline{F}_{ab} \ \sin(\alpha\varphi) \tag{36}$$

In (36) $\overline{F}_{ab}$ is an antisymmetric tensor field in space-time which vanishes outside of some tube in space-time. This ensures that, in terms of the "space and time" point of view, the F_{ab} of Eqn. (36) represents a packet of electromagnetic waves which moves through space. (The question we must answer, from Maxwell's equations, is "How does it move.") The φ in (36) is some scalar field in space-time. Finally the α in (36) is just a number. We are interested in (36) in the limit of large α. When α is large, the slightest change in φ causes $\alpha\varphi$ to go through many multiples of 2π. Hence, $\sin(\alpha\varphi)$ oscillates between $+1$ and -1 many times. That is, the limit of (36) for large α is the limit of high frequency and short wavelength. i.e., the optical limit.

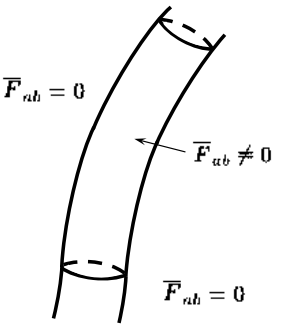
We ask: what are the conditions of φ and $\overline{F}_{ab}$ in order that (36) satisfy Maxwell's equations, (25) and (26), in the limit of large α? (We set $J^a = 0$, for we are interested in free wave solutions.) Substituting (36) into (25) and (26), we obtain

$$\sin(\alpha\varphi)\nabla^{[a}\overline{F}^{bc]} + \alpha\cos(\alpha\varphi)(\nabla^{[a}\varphi)\overline{F}^{bc]} = 0$$
$$\sin(\alpha\varphi)\nabla_b\overline{F}^{ab} + \alpha\cos(\alpha\varphi)(\nabla_b\varphi)\overline{F}^{ab} = 0,$$

respectively. For α large, the first term in each equation above is negligible compared to the second, Hence, setting $k^a = \nabla^a\varphi$, we have a solution in the optical limit provided

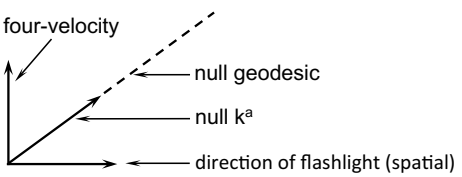
$$k^{[a}\overline{F}^{bc]} = 0 \tag{37}$$

$$k_b\overline{F}^{ab} = 0 \tag{38}$$



Contracting (37) with k_a, using (38), we obtain $\frac{1}{3}k_a k^a F^{bc} = 0$. Hence, $k_a k^a = 0$, i.e., k^a is a null vector field. But now $k^b \nabla_b k_a = k^b \nabla_b \nabla_a \varphi = k^b \nabla_a \nabla_b \varphi = k^b \nabla_a k_b = \frac{1}{2}\nabla_a (k^b k_b) = 0$. Hence, k^b is tangent to null geodesics. All that remains is to interpret these results.

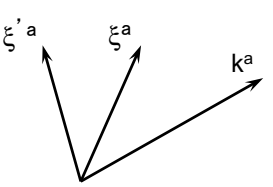
First note that the surfaces of constant phase for the wave are the surface of constant φ. Hence, $k^a = \nabla^a \varphi$, which is normal to these surfaces of constant phase, is the direction of propagation of the wave in space-time. We conclude that *light rays are null geodesics in space-time*. (More precisely, a packet of electromagnetic waves, in the optical limit, moves on a null geodesic.) Thus, light goes, not along any old null curve, but along a particular type of null curve. (This is ok. If you want to send out a pulse of light with a flashlight, you get to choose the event at which the light is sent, and the direction (in space) in which the light is emitted. Given a point in space-time, a four-velocity there, and a direction orthogonal to that four-velocity, there is precisely one null direction whose projection orthogonal to the four-velocity is the given (spatial) direction. But a null direction at a point of space-time uniquely determines a corresponding null geodesic.)



A direct physical interpretation of k^a can be obtained by introducing an observer. Set $\alpha = 1$ in (36). (That is, choose α large enough so the optical limit is valid, then rescale φ by a constant factor to get $\alpha = 1$). Let our observer have four-velocity ξ^a. Now, $\xi^a k_a = \xi^a \nabla_a \varphi$ is the rate of change (with time) of the phase of the wave as seen by our observer. That is to say, $\omega = -\xi^a k_a$ is the apparent (angular) frequency of the wave, as seen by our observer. Set $K_a = h_a{}^b k_b$, a spatial vector according to our observer. Then if v^a is a unit vector in the direction of K^a, $v^a K_a = v^a \nabla_a \varphi$ is the rate of change of the phase of the wave with respect to distance, in the spatial direction v^a (according to this observer). When the phase has gone through 2π, then one has passed form one crest of the wave to the next. Hence, $\lambda = 2\pi / v^a K_a$ is the wavelength of the light, as seen by this observer. This can be simplified by recalling that v^a is a unit vector in the direction of K_a. Hence,

$$\lambda = \frac{2\pi}{v^a K_a} = \frac{2\pi}{\dfrac{K^a K_a}{(K^m K_m)^{1/2}}} = 2\pi (K^m K_m)^{-1/2}$$
$$= 2\pi (h^{ab} k_a k_b)^{-1/2} = 2\pi [(g^{ab} + \xi^a \xi^b) k_a k_b]^{-1/2} = -2\pi (\xi^a k_a)^{-1}$$

Note that the angular frequency and wavelength are related by $\lambda = 2\pi\omega^{-1}$, as we would expect for a wave traveling (according to our observer) at the speed of light.



All the usual formulae for Doppler shifts, aberration, etc. follow immediately and easily from the remarks above. As an example, we work out the formula for the Doppler effect. Let two observers pass through the event p in space-time, and let their four-velocities be ξ^a and ξ'^a. A light ray (with its k^a) also comes to the event p. The two observers see it, and assign to it frequencies $\omega = -\xi^a k_a$ and $\omega' = -\xi'^a k_a$. How are these

frequencies related? It depends, of course, on the relationship between ξ^a, ξ'^a and k_a. Let us suppose that these three vectors are linearly dependent. (Physical interpretation: each observer sees the other moving in the direction, or opposite the direction, of the light ray.) Thus, we are doing what is called the longitudinal Doppler shift. (For the transverse Doppler shift, one would require that the projection of ξ'^a orthogonal to ξ^a is orthogonal to the projection of k^a orthogonal to ξ^a. That is, one would require that the ξ-observer see the light ray and the ξ'-observer moving in orthogonal directions.) Since ξ^a, ξ'^a and k^a are dependent, k^a is a linear combination of ξ^a, ξ'^a : $k^a = m\xi^a + m'\xi'^a$. Setting $k^a k_a = 0$, and solving for m and m' (we actually care about their ratio), we have $k^a = s[-\xi^a(\xi^m \xi'_m - \sqrt{(\xi^m \xi'_m)^2 - 1}\,) - \xi'^a]$ where s is number. Hence,

$$\omega = -\xi_a k^a = s\sqrt{(\xi^m \xi'_m)^2 - 1}$$
$$\omega' = -\xi'_a k^a = s[(\xi^m \xi'_m)^2 - 1 - (\xi^m \xi'_m)\sqrt{(\xi^m \xi'_m)^2 - 1}]$$

So,

$$\frac{\omega'}{\omega} = -\xi^m \xi'_m + \sqrt{(\xi^m \xi'_m)^2 - 1}$$

This is the required formula. It is convenient, however, to express it in terms of apparent relative speed between the observers. (See Eqn. (10), writing ξ'^a for ξ^a.) Substituting (10), we obtain, finally,

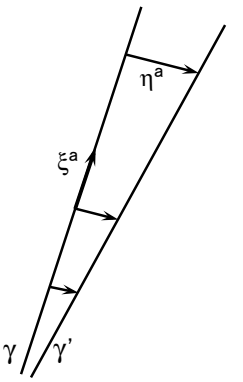
$$\frac{\omega'}{\omega} = \sqrt{\frac{1+v}{1-v}}$$

this will be recognized as the familiar formula for the longitudinal Doppler shift.

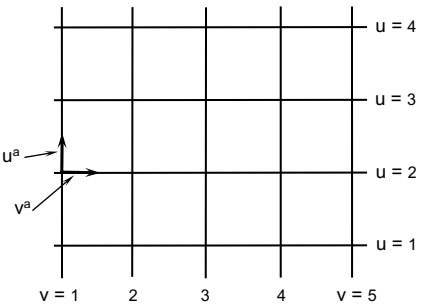
As far as I am aware, Lorentz transformations are never a good way to obtain the standard formulae of special relativity. Instead, one can set up the problem in terms of the relevant vectors and tensors, work out the inner product which yield what one wants, and solve for what one wants in terms of what one is given. Furthermore, these formulae are, in my opinion, almost never useful. The tensors are simple and clear. The formulae which relate these tensors to the experience of an observer should, perhaps, be seen once, to make it clear how they can be obtained. Beyond that, one can happily live with just the tensors.

18. Geodesic Deviation

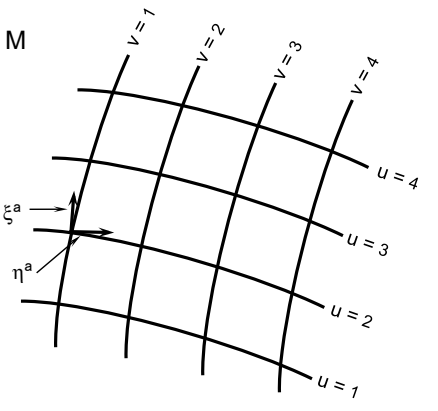
Let γ be a straight line in Euclidean 3-space, with unit tangent vector ξ^a. Let γ' be a nearby straight line. At each point of γ, let η^a be the connecting vector from that point, orthogonal to γ, to the corresponding point of γ'. It should be clear geometrically that $\xi^n\xi^m\nabla_n\nabla_m\eta^a = 0$, i.e., the second derivative of the connecting vector along γ vanishes. Now suppose we replace Euclidean 3-place by space-time, and the straight line by geodesics. Then the second derivative of the connecting vector along the geodesic is, rather then zero, an expression involving the curvature of space-time. We now derive this expression. It gives a pretty, intuitive, "geometrical" picture of what curvature in space-time does.



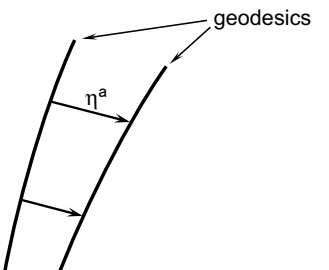
It is convenient to replace the intuitive discussion involving "nearby curves" above by a one-parameter family of curves. We thus begin as follows. Let $S = \mathbb{R}^2$ (the plane) with coordinates (u, v). Introduce vector fields u^a and v^a on S such that $u^a\nabla_a u = v^a\nabla_a v = 1$, $u^a\nabla_a v = v^a\nabla_a u = 0$. (It is not difficult to check that these properties determine u^a and v^a uniquely.) Thus, we have a two-dimensional "grid".



The idea is to "insert" this grid into space-time, to obtain our one-parameter family of curves (v will label the various curves, and u will be a parameter along each curve). This we do as follows: let $\psi : S \to M$ be a smooth mapping. Then, fixing v, $\psi(u, v)$, as u varies, describes a curve in space-time, where u serves as a parameter along this curve. Thus, we obtain a one-parameter family of curves (labeled by v). Denote by ξ^a and η^a, respectively, the images of u^a and v^a under this mapping. That is, set $\xi^a = \vec{\psi}\,u^a$, $\eta^a = \vec{\psi}\,v^a$. Evidently, ξ^a is just the tangent vectors to our curves, while η^a is the connecting vector between successive curves. Thus, by introducing a family of curves, and taking the "derivative of position in space-time with respect to the curve (i.e., with respect to this parameter)", we obtain, in a clean way a vector η^a which represents, intuitively, the "connecting vector between curves".



We are interested, not in arbitrary one-parameter families of curves, but rather in one-parameter families of timelike geodesics. Hence, we set $\xi^a\xi_a = -1$, and $\xi^b\nabla_b\xi^a = 0$. Thus, u measures proper time along each timelike geodesics. Fur-



thermore, it follows from our construction that $\xi^m \nabla_m \eta^a = \eta^m \nabla_m \xi^a$. (Proof: This says that the Lie derivatives of η^a in the ξ-direction vanishes. But, clearly, the Lie derivative of v^a in the u-direction vanishes. The result now follows from the fact that Lie derivatives are carried forward by smooth mappings.) Geometrically, $\xi^m \nabla_m \eta^a = \eta^m \nabla_m \xi^a$ states that, if the "foot" of the infinitesimal displacement sits on one geodesic, and the "tip of the arrow" sits on another, then, as η^a is introduced at successive points along the geodesics, it continues to point from the same geodesic. In other words, this is precisely the condition that η^a really represent the "connecting vector" from one geodesics to a neighboring one.

Having set up what is to be calculated, the calculation itself is easy:

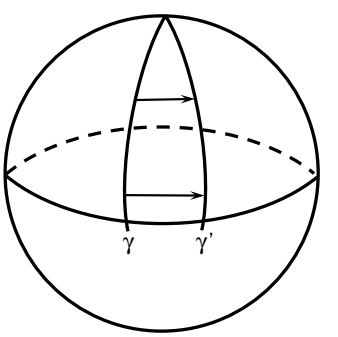
$$\begin{aligned}
&\xi^m \xi^n \nabla_m \nabla_n \eta_a = \xi^m \nabla_m (\xi^n \nabla_n \eta_a) - (\xi^m \nabla_m \xi^n) \nabla_n \eta_a \\
&= \xi^m \nabla_m (\xi^n \nabla_n \eta_a) \\
&= \xi^m \nabla_m (\eta^n \nabla_n \xi_a) \\
&= (\xi^m \nabla_m \eta^n) \nabla_n \xi_a + \xi^m \eta^n \nabla_m \nabla_n \xi_a \\
&= (\xi^m \nabla_m \eta^n) \nabla_n \xi_a + \xi^m \eta^n \nabla_n \nabla_m \xi_a + \xi^m \eta^n R_{mnac}\, \xi^c \\
&= (\xi^m \nabla_m \eta^n) \nabla_n \xi_a + \eta^n \nabla_n (\xi^m \nabla_m \xi_a) - (\eta^n \nabla_n \xi^m) \nabla_m \xi_a \\
&+ \xi^m \eta^n R_{mnac}\, \xi^c \\
&= -(R_{manb}\, \xi^m \xi^n) \eta^b
\end{aligned}$$

where, in the first step, we have differentiated by parts; in the second, we have used $\xi^m \nabla_m \xi^n = 0$; in the third, we have used $\xi^n \nabla_n \eta^a = \eta^n \nabla_n \xi^a$; in the fourth, we have expanded using the Leibniz rule: in the fifth, we have used the definition of the Riemann tensor, $\nabla_m \nabla_n \xi_a = \nabla_n \nabla_m \xi_a + R_{mnac}\, \xi^c$; in the sixth, we have differentiated the middle term in the preceding expression by parts. Finally, in the last step, we have used the fact that the first and third terms in the preceding expression cancel, and that the second term is zero by geodesic-ness. We have:

$$\xi^m \xi^n \nabla_m \nabla_n \eta_a = -(R_{manb}\, \xi^m \xi^n) \eta^b \tag{39}$$

This is called the *equation of geodesic deviation*. Note that, for flat space ($R_{abcd} = 0$), we recover the observation at the beginning of this section.

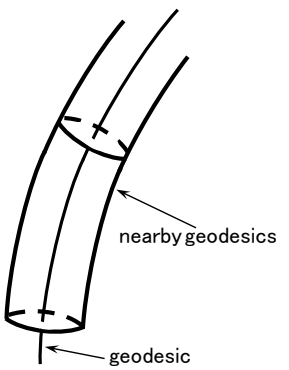
Why do we wind up with an equation such as (39)? First note that curvature could not affect a single geodesic, for there would be nothing to compare it with (i.e., that there is no natural "thing which is like a geodesic, but uninfluenced by a curvature") In light of this, curvature does the best it can: it manifests itself when we compare one geodesic with a nearby one. Example: geodesic (great circles) on a 2-sphere. The curvature "draws nearby geodesics together". Why is it that the second derivative of η^a (along the geodesic) appears on the left in (39), rather than, say, the first derivative? The point here is that $\xi^m \nabla_m \eta^a$ is freely at our disposal. We can "start out" our two nearby geodesic running parallel to each other ($\xi^m \nabla_m \eta^a = 0$ initially), separating from each other ($\xi^m \nabla_m \eta^a \propto \eta^a$ initially), moving "sideways relative to each other" ($\xi^m \nabla_m \eta^a$ orthogonal to η^a initially). However, after we fix how the two nearby geodesics



begin, the geodesic equations take over. In particular, they determine the second derivative of the connecting vector along the geodesic. Hence, the second derivative is the first derivative not at our disposal, and hence the first derivative, which could possibly appear in (39). (Expressions for, e.g., $\xi^m\xi^n\xi^p\nabla_m\nabla_n\nabla_p\eta^a$ are obtained immediately from (39).)

Setting $K_{ab} = -R_{manb}\ \xi^m\xi^n$, Eqn. (39) becomes $\xi^m\xi^n\nabla_m\nabla_n\eta_a = K_{ab}\eta^b$. Note that K_{ab} is a symmetric, spatial tensor according to the observer with four-velocity ξ^a. ($K_{ab} = K_{(ab)}$. Proof: $R_{manb} = R_{nbma}$; $\xi^b K_{ab} = 0$. Proof: Since $R_{manb} = R_{ma[nb]}$, $\xi^m\xi^n\xi^b R_{manb} = 0$.) Hence, if η^a and $\xi^m\nabla_m\eta^a$ are initially orthogonal to ξ^m, this remains true along the geodesic. We can, without loss of generality, impose this orthogonality; and do so. There follows an intuitive discussion of the geometry of this K_{ab}.

Let, α, λ^b be an eigenvalue-eigenvector of K_{ab}, i.e., let $K_{ab}\lambda^b = \alpha\lambda_a$. Choose $\eta^a = \lambda^a$, and $\xi^m\nabla_m\eta^a = 0$ initially, where we write a dot for $\xi^m\nabla_m$. Then (39) gives $\ddot{\eta}^a = \alpha\eta^a$. In other words, this particular η^a merely lengthens or shortens (α positive or negative, respectively) as we move along the geodesic. Now consider the three independent eigenvalues-eigenvectors of K_{ab}. Each lengthens or shortens, depending on its eigenvalue. But a general η^a can be written as a linear combination of these eigenvectors. Thus, we have a complete description of the action of (39). We set things up as follows. At some point on our basic geodesic, begin nearby geodesics all equidistant from the given geodesic, and running parallel to it (i.e., all have $\xi^m\nabla_m\eta^a = 0$) Then, to second order as we move away from this point along our basic geodesic (second order cause in (39)), the small 2-sphere described by the neighboring geodesics begins distorting into an ellipsoid. The axes of the ellipsoid are the eigenvectors of K_{ab}. Thus, if one eigenvalue is positive and two negative, we obtain a "cigar", while one negative and two positive yield a "pancake". Finally, note that the rate of change of the volume of the sphere is the sum of the eigenvalues, i.e., $K^m{}_m$.



We now repeat this description physically. You, the observer, are in free fall (i.e., a geodesic). At some instant, you release a lot of particles, all at rest relative to you, and all lying in a sphere of radius one foot from you. The derivative of the shape of the sphere (with time) is zero, since the particles were initially at rest. However, the second derivative with time results in a distortion of the sphere into an ellipsoid. The axes of the ellipsoid are the eigenvectors of K_{ab}. The second time rate of change of the volume of the sphere (as it goes into an ellipsoid) is $K^m{}_m$.

Clearly, the Riemann tensor is determined completely by its effects in causing geodesic deviation, i.e., by Eqn. (39).

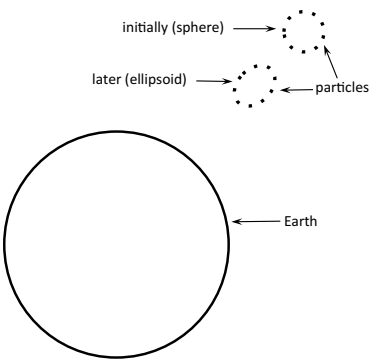
19. Einstein's Equation

We shall now write down Einstein's equation – the equation analogous to $\nabla^2\varphi = 4\pi G\rho$ in Newtonian theory.

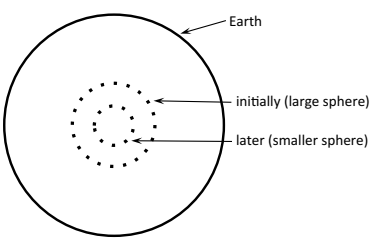
We begin by asking the question: what is the relative behavior of nearby particles in Newtonian theory, i.e., what is "Newtonian geodesic deviation" like? Recall that the motion of a particle in Newtonian theory is given by $x^a(t)$, the position of the particle in Euclidean space as a function of time t. The equation of motion of such a particle (if uninfluenced by other forces) is $\ddot{x}^a = -\nabla^a\varphi$. Now suppose we have two nearby free particles, with positions x^a and $x^a + \delta x^a$. Then, to first order in δx^a, we have

$$\begin{aligned}(\delta x^a)^{\cdot\cdot} &= (x^a + \delta x^a)^{\cdot\cdot} - (x^a)^{\cdot\cdot} = -\nabla^a\varphi(x + \delta x) + \nabla^a\varphi(x) \\ &= -(\nabla^a\nabla^b\varphi)\delta x_b. \end{aligned} \tag{40}$$

Eqn. (40) gives the effect of ordinary Newtonian gravitation on nearby particles. Suppose, for example, that we begin with a swarm of particles initially at rest with respect to the Earth, and above the Earth, where these particles form a 2-sphere. The particles are released. Since the Earth attracts the particles closer to it more than those farther away, a bit later the sphere becomes an ellipsoid, with the axis pointing toward the Earth. If, on the other hand, we began with particles forming a sphere about the center of the Earth, then a bit later they would still describe a sphere, but a smaller one (all particles fall toward the center of the Earth).



We regard Eqn. (40) as closely analogous to Eqn. (39). Here, x^a corresponds to η^a a dot corresponds to $\xi^m\nabla_m$ ("time-derivatives"), whence $\nabla_a\nabla_b\varphi$ is analogous to $R_{manb}\xi^m\xi^n$. Thus, *because of the geodesic behavior of free particles in space-time, the quantity $R_{manb}\xi^m\xi^n$ has the same effect on nearby free particles in space-time as has $\nabla_a\nabla_b\varphi$ on nearby free particles in Newtonian theory.*



The interpretation of $\nabla^2\varphi = 4\pi G\rho$ in Newtonian theory should now be clear. It says that the second time derivative of the volume of a small sphere of free particles is given by the density of matter within the sphere. (Note that $\nabla^2\varphi$ is the sum of eigenvalues of $\nabla_a\nabla_b\varphi$.) Thus, in the figure on the preceding page, there is no matter inside the sphere, of particles. Hence, the particles begin to form an ellipsoid, but with the same volume as the original sphere. In the figure on this page, the decrease in the volume of the sphere with time is due to the presence of matter (the Earth) inside the sphere of particles. Fix a particle of mass m in Newtonian theory. Consider particles, begun at rest, one foot from this one. Then $\nabla^2\varphi = 4\pi G\rho$ states that, on the average (over directions in which the second particle lies from the first) the two particles fall closer together by Newton's law of gravitation. Because of the presence of other masses outside this

system, some of the particles (on the one-foot sphere) may fall toward the central particle more quickly; others more slowly. But $\nabla^2\varphi = 4\pi G\rho$ fixes the average.

Since $R_{manb}\,\xi^m\xi^n$ is analogous to the Newtonian $\nabla_a\nabla_b\varphi$, and since we have $\nabla^2\varphi = 4\pi G\rho$ in Newtonian theory, one would naturally set $R_{ma}\,\xi^m\xi^n = 4\pi G\rho$ as the equation for the creation of curvature by matter in general relativity. This will ensure that each free observer sees nearby free particles falling in toward him, on the average, by an amount determined by his ρ. In other words, this will ensure that our observer, locally, sees the validity of Newtonian gravitation.

Unfortunately, $R_{ma}\,\xi^m\xi^n = 4\pi G\rho$ still contains this quantity ρ, the Newtonian mass density. By what should it be replaced is space-time? Unfortunately, there are two choices. One could set $\rho = T_{mn}\,\xi^m\xi^n$, the local mass density seen by an observer with four-velocity ξ^m, where T_{mn} is the total stress-energy. There is a second expression, based on the fact that, in the Newtonian limit the mass density dominates (i.e., pressures, momentum densities, etc. are all small). That is, in the Newtonian limit, $T_{mn} = \rho\xi_m\xi_n$. Hence, we could equally well set $\rho = -T$ as the space-time object to replace the Newtonian ρ.

More generally, we could consider some linear combination – $\rho = rT_{mn}\,\xi^m\xi^n - (1-r)T$, where r is a number, Then our generalization of $\nabla^2\varphi = 4\pi G\rho$ would be

$$R_{mn}\,\xi^m\xi^n = 4\pi G[rT_{mn} + (1-r)Tg_{mn}]\xi^m\xi^n. \tag{41}$$

Since this must hold for all unit timelike ξ^m, we have Einstein's equation:

$$R_{ab} = 4\pi G[rT_{ab} + (1-r)Tg_{ab}]. \tag{42}$$

In short, we obtain a one-parameter family (labeled by r) of generalization of the Newtonian law.

What value shall we assign to r? We first rewrite Eqn. (42). Contracting over "a" and "b", solving for T in terms of R, and substituting into (42), we have

$$R_{ab} - \frac{r-1}{3r-4}Rg_{ab} = 4\pi\,G\,r\,T_{ab} \tag{43}$$

Recall that the total stress-energy is conserved: $\nabla_b T^{ab} = 0$. Furthermore, Bianchi's identity yields $\nabla_m(R^{ab} - \frac{1}{2}Rg^{ab}) = 0$. The natural choice for r is that wish makes these equations consequences of each other. That is, we are led to set $r = 2$, so both sides of (43) are divergence-free, Thus, we obtain *Einstein's equation*:

$$R_{ab} - \frac{1}{2}Rg_{ab} = 8\pi\,G\,T_{ab} \tag{44}$$

(That both sides of (44) be conserved in actually required, as we shall see later, for the existence of an initial-value formulation of the theory.)

We regard, the Einstein equation, (44), as requiring that the curvature of space-time be so related to the matter distribution that free particles, responding to the curvature, behave as do free particles responding to the Newtonian potential behave in Newtonian theory. To put matters another way, it is clear that everything we have done has the correct special relativity limit (when $R_{abcd} = 0$). Eqn. (44) ensures that we also have the correct limit of Newtonian gravitation.

Note: $G = 6.67 \times 10^{-8}$ dyne cm^2/g^2 $= 2.2 \times 10^{-39}$ sec/g.

20. Einstein's Equation: General Remarks

Consider Maxwell's Equations, Eqns. (25) and (26). It is often the case in practice that one adopts the following attitude toward these equations. The charge-current J^a is given, and one wishes to solve (25) and (26) for the electromagnetic field, F_{ab}. This would be the situation, for example, if we were given the current in a system of wires (i.e., given J^a), and wished to find the resulting the electromagnetic field. In other situations, however, the charge-current is strongly influenced by F_{ab}, and so no such procedure is available for solving (25) and (26). This would be the case, for example if we were dealing with a cloud of charged particles, so the value of F_{ab} determines the motion of the cloud, and hence J^a. Similar remarks apply to $\nabla^2\varphi = 4\pi G\rho$ in Newtonian gravitation. To find the gravitational field of a given object (i.e., an elephant), one solves for φ, with ρ given. To determine the motion of the planets, φ affects what ρ is, so no such simple procedure is available.

How does Einstein's equation fit into this framework? *It is unreasonable physically to regard T_{ab} as given* (on a manifold), *with Einstein's equation (44) to be solved for the metric g_{ab}*. The reason is that T_{ab} itself is a quantity which refers, not only to "matter", but also to "geometry". Consider, for example, the stress-energy of a fluid, $T_{ab} = (\rho + p)\xi_a\xi_b + pg_{ab}$. Now, ξ^a, the four-velocity of the fluid, is unit ($\xi^a\xi^b g_{ab} = -1$); ρ is the mass density in grams/sec^3; the metric occurs in the second term. Everywhere, we see the metric, directly or indirectly, in the stress-energy. One should not regard "specifying T_{ab}" as the same as "specifying what the matter is doing". The relation between T_{ab} as a tensor field and "what the matter is doing" involves the metric. It appears that it is simply impossible to make any reasonable description of matter without the notions of space and time provided by the metric.

In light of the remarks above, one might be tempted to proceed in the other direction. We give g_{ab} and "solve" (trivially) (44) for T_{ab}. One could indeed regard the result as a "solution" of Einstein's equation. The difficulty is that the resulting T_{ab} will not, without extraordinary luck, resemble any kind of matter of interest.

As far as I am aware, there exists no technique for generating all physically interesting solutions of Einstein's equation. The collection of known exact solution is relatively small, and this makes every exact solution potentially valuable (either because it might provide a prediction of general relativity, or – often as important – because it might provide insight into the structure of the theory itself). Those solutions which are available are obtained by special techniques, e.g., assumptions of symmetry, or assumptions of simple structure for the Riemann tensor. We shall later introduce some techniques for obtaining solutions of Einstein's equation.

Recall the discussion of Sect. 16. We can regard gravitation as just another "type of matter". The corresponding tensor field is the metric, g_{ab}. The equation on this field is Einstein's equation. Of course, the metric needn't have a stress-energy: that's the mechanism by which other types of matter influence the metric. The situation is similar to that of electromagnetism. All "types of charge" end up producing a charge-current density, J^a. If several "types of charge" are available, we add their contributions to obtain the total charge-current J^a. This acts as a source for the electromagnetic field. This total charge-current must have vanishing divergence (in order that there exist solutions of (25), and (26)) If in fact the

"types of charge" are able to exchange charge between them, then the individual J^a's for the individual types of charge will not be conserved: only the total. The electromagnetic field is not obligated to come up with a charge-current for it. (The analogy: J^a for T_{ab}; vanishing divergence for conserved; F_{ab} for g_{ab}.)

What is so special, from the point of view above, that sets gravitation apart from other fields? It is the universality of gravitation. The metric enters into everybody else's stress-energy – it effects all types of matter. Every type of matter produces a stress-energy, and hence affects the geometry of space-time. (In the electromagnetic case, by contrast, some things are uncharged, and so do not interact with electromagnetism.) Hence, all possible measuring instruments are affected by g_{ab}. Because of this universality, any attempt to measure features of the geometry of space-time will be influenced by gravitation. Hence, gravitation gets inseparably intertwined with geometry. Thus, the metric, g_{ab} has two roles: as giving geometry of space-time, and as a gravitational potential. (Think of g_{ab} as analogous to Newtonian φ. Then $\nabla_a \nabla_b \varphi$ is analogous to the "second derivative" of the metric, i.e., to the Riemann tensor. That's what we want from Sect. 19.)

21. The Friedmann Solution

The information presently available about the structure of our universe is rather meagre. However, the following two features do appear to be comparatively well-established. *The dominant matter in the universe* (with regard to the dynamics of the universe) *is the matter in the galaxies, which may be regarded as dust.* Essentially all the matter of which we are aware in the universe is that in the galaxies. There are some 10^{11} galaxies, enough so that we might expect to be able to regard each as a "particle of dust". There are, of course, other things in the universe (e.g., electromagnetic radiation), but these contribute to the total stress-energy a negligible amount. *The universe is spatially isotropic, as seen from any galaxy.* (The mathematical formulation of "spatially isotropic" will come later.) First note that the universe is not isotropic, in its details as seen from Earth. For example, we see nearby galaxies in some directions and not in others. But these are local features. In the large (i.e., over distances large compared with the distance between galaxies), the universe does indeed appear isotropic from the Earth. For example, we see about as many galaxies in one region of the sky as in another, on the average. From bitter previous experience, we do not wish to regard the location of the Earth in the universe as somehow special. Hence, we are led to assume isotropy for everybody.

It is an easy matter to obtain the most general solution of Einstein's equation compatible with the features above. If general relativity is correct, and if the features above are indeed valid for our universe, then one of these solutions will describe the overall behavior of our universe. We now obtain these, the Friedmann solutions.

The matter is dust. Hence, we have the density ρ of the dust, the four-velocity ξ^a of the dust, and

$$T_{ab} = \rho \xi_a \xi_b \tag{45}$$

Next, consider $\nabla_a \xi_b$. The antisymmetric part, $\nabla_{[a}\xi_{b]}$ is an antisymetric tensor field in space-time. Just as we did for the electromagnetic field in Sect.13. this $\nabla_{[a}\xi_{b]}$ can be decomposed (by observer with four-velocity ξ^a, i.e., by us) into two spatial vectors. But this would violate isotropy! If a spatial vector could be obtained from the velocity field of the galaxies and the metric, then some directions in space would be preferred over others. Since we wish to impose isotropy – since we wish that no spatial directions are preferred – we must have $\nabla_{[a}\xi_{b]} = 0$. Thus, $\nabla_a \xi_b$ must be symmetric in "a" and "b". Next, note that $\xi^b \nabla_a \xi_b = 0$ (since ξ^b is unit), and $\xi^a \nabla_a \xi_b = 0$ (geodesic motion of dust particles). Hence, $\nabla_a \xi_b$ is a (symmetric) spatial tensor (according to observer with four-velocity ξ^a). Since $\nabla_a \xi_b$ lives in an odd number of dimensions (i.e., since its characteristic equation is a cubic), this tensor must have some eigenvector, λ^a. That is, let λ^a be spatial, with $\lambda^b(\nabla_a \xi_b) = \alpha \lambda_a$. This spatial vector would be preferred (violating isotropy) unless $\mu^b(\nabla_a \xi_b) = \alpha \mu_a$ for all spatial μ_b. Thus,

$$\nabla_a \xi_b = \alpha h_{ab} \tag{46}$$

for some scalar field α, where, as usual, we set $h_{ab} = g_{ab} + \xi_a \xi_b$. To interpret this α physically, contract (46): $\nabla^a \xi_a = 3\alpha$. That is, the divergence of ξ^a is 3α, so α represents the "rate of expansion of the universe". We shall call α the *Hubble field* (for reasons which will emerge shortly).

To impose Einstein's equation, we need the Ricci tensor, which is obtained from the Riemann tensor, which is obtained by commuting derivatives. The obvious thing on which to commute derivatives is ξ_c. We have

$$\begin{aligned}\nabla_a(\nabla_b\xi_c) &= (\nabla_a\alpha)h_{bc} + \alpha\nabla_a h_{bc} \\ &= -\dot\alpha\xi_a h_{bc} + \alpha\nabla_a h_{bc} = -\dot\alpha\xi_a h_{bc} + \alpha\nabla_a(g_{bc}+\xi_b\xi_c) \\ &= -\dot\alpha\xi_a h_{bc} + 2\alpha(\nabla_a\xi_{(b})\xi_{c)} \\ &= -\dot\alpha\xi_a h_{bc} + 2\alpha^2 h_{a(b}\xi_{c)}\end{aligned} \tag{47}$$

where, for the first step, we have substitute (46); for the second, used $\nabla_\alpha = -\dot\alpha\xi_a$, where a dot denotes $\xi^m\nabla_m$ (by isotropy, $\nabla_a\alpha$ cannot have a non vanishing component orthogonal to ξ_a. Hence, $\nabla_a\alpha$ a multiple of ξ_a; for the third, used $h_{bc} = g_{bc}+\xi_b\xi_c$; for the fourth, expanded using the Leibniz rule; and, for the fifth, again used (46). Now antisymmetrize (47) over "a" and "b", using $\nabla_{[a}\nabla_{b]}\xi_c = \frac{1}{2}R_{abcd}\,\xi^d$:

$$R_{abcd}\,\xi^d = -2(\dot\alpha+\alpha^2)\xi_{[a}h_{b]c} \tag{48}$$

Thus, we obtain immediately from (46) a simple expression for "most" of the Riemann tensor. Isotropy is a very strong condition, which produces a considerable simplification.

To use Einstein's equation, we need the Ricci tensor. Hence, we contract (48) over "a" and "c" (after raising one of the indices, of course. We won't normally state that explicitly.) Hence, from (48):

$$R_{bd}\,\xi^d = -3(\dot\alpha+\alpha^2)\xi_b \tag{49}$$

We next use Einstein's equation, (44) but in the form $R_{ab} = 8\pi\,G(T_{ab} - \frac{1}{2}Tg_{ab})$, where $T = T^m{}_m$. (To obtain this form, contract (44) to obtain $R = -4\pi\,G\,T$. Substitute this for R in (44).) Hence, from (45),

$$R_{ab} = 8\pi\,G(\rho\xi_a\xi_b + \frac{1}{2}\rho g_{ab}) \tag{50}$$

Finally, substitute R_{ab} from (50) into (49) to obtain

$$\dot\alpha + \alpha^2 = -\frac{4}{3}\pi\,G\,\rho \tag{51}$$

Eqn. (51) has a simple physical interpretation. The $\dot\alpha$ is "the rate of change with time of the rate of expansion of the universe". Hence, $\dot\alpha$ is an "acceleration". Ignoring factors and the α^2 term, Eqn. (51) says that "the universe has negative acceleration, with magnitude ρ". In other words, "the attraction of all those galaxies for each other tends to slow the rate of expansion of the universe". This is (part of) the effect of the dust on the geometry of space-time required by Einstein's equation.

Eqn. (51) alone cannot be solved. We have $\alpha(t)$ and $\rho(t)$ (where t is proper time along a galaxy world-line). (Clearly, we expect the mass density of galaxies, ρ, to change with time as the universe expands.) To obtain a second equation, we use conservation of dust: $0 = \nabla_a(\rho\xi^a) = \rho\nabla_a\xi^a + \xi^a\nabla_a\rho$. Substituting (46),

$$\dot\rho + 3\alpha\rho = 0 \tag{52}$$

Eqn. (52) is also clear physically. It states that "the mass density of the dust decreases at a rate proportional to the rate of expansion of the universe" – exactly what we would expect from conservation of dust particles.

We now wish to solve the simultaneous Eqns. (51) and (52). Since α is the "rate of expansion of the universe", it is convenient to introduce a function $r(t)$ such that α is the "rate of expansion of r". Set $\alpha = \dot{r}/r$. (Remark on the physical interpretation of r: Let η^a be a spatial vector along a galaxy world-line, and suppose η^a joins this world-line to the world-line of nearby galaxy. Then, setting $r = (\eta^a \eta_a)^{\frac{1}{2}}$, we have $\dot{r}/r = \alpha$. (Hint: Use $\pounds_\xi \eta^a = 0$.) Thus, r could be the "spatial distance of the Andromeda from us".) Setting $\alpha = \dot{r}/r$ in (52), we have $3\dot{r}\rho + r\dot{\rho} = 0$, or $(r^3\rho) = \text{const.}$ (What does $r^3\rho$ const. mean physically?) We make the following choice of this constant (the constant is at our disposal because $r(t)$ is only defined up to a constant factor):

$$\frac{4}{3}\pi G \rho = r^{-3} \tag{53}$$

Substituting (53) and $\alpha = \dot{r}/r$ into (51), we easily obtain

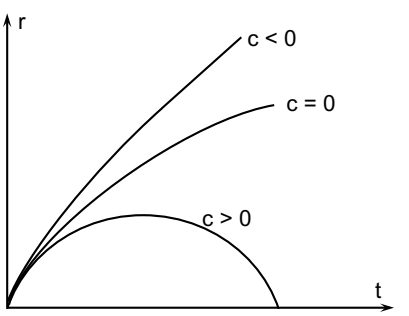
$$\ddot{r} = -r^{-2} \tag{54}$$

Eqn. (54) also has a simple physical interpretation. Think of $r(t)$ as a "characteristic size of the universe". Then $\ddot{r}$ is the "acceleration". Hence, from (54), we regard $-r^{-2}$ as an "effective (attractive, inverse r-squared) force". In fact, Eqn. (54) is precisely the equation describing a brick thrown radially from the Earth (where r = distance of brick from Earth's center). The acceleration of the brick ($\ddot{r}$) equals the force on the brick ($-r^{-2}$). It is now clear, from our intuitive understanding of bricks, what the solutions of (54) (and, hence, of (51) and (52)) will look like. Bricks which are thrown upward with a small initial speed soon fall back. Those which have a large initial speed escape from the Earth. There is a critical speed between these extremes. These bricks escape, but with a speed which diminishes toward zero in the limit as the brick gets far away from the Earth.

The analogy above also suggests how to solve (54): use conservation of energy. Multiplying (54) by $\dot{r}$, we have $2r^{-1} - (\dot{r})^2 = c = \text{const.}$ (For the brick, this is the binding energy. The brick returns precisely when c is positive.) Thus, $\dot{r} = \pm(2r^{-1} - c)^{\frac{1}{2}}$, which is easily solved with a table of integrals. It is simplest to express the solution in parametric form:

$$\begin{array}{lll} c > 0 & r = c^{-1}(c - \cos x) & t = c^{-3/2}(x - \sin x) \\ c = 0 & r = \frac{1}{2}x^2 & t = \frac{1}{6}x^3 \\ c < 0 & r = (-c)^{-1}(\cosh x - 1) & t = (-c)^{-3/2}(\sinh x - x) \end{array} \tag{55}$$

where x is a parameter. These functions are graphed below.



These graphs are, of course, exactly what we would expect from the brick analogy. Of course, given $r(t)$, one determines immediately $\alpha(t)$ and $\rho(t)$.

The discussion above refers to the "dynamics" of the Friedmann solutions. We now describe their "spatial geometry". This discussion is based on the following remark: if $W_{a\ldots c}$ is any spatial tensor (according to ξ^a), then $\nabla_m W_{a\ldots c}$, with all its indices projected orthogonal to ξ^a, is the spatial derivative of this tensor. (Proof: $\nabla_a h_{bc}$, with all

its indices projected, gives zero. See definition of covariant derivative associated with a metric.) Let l_a be any spatial vector field in our space-time. Then

$$\frac{1}{2}\mathscr{R}_{abcd}\, l^d = h_{[a}{}^r h_{b]}{}^s h_c{}^t\, \nabla_r [h_s{}^m h_t{}^n \nabla_m l_n] \tag{56}$$

by definition of the spatial Riemann tensor, $\mathscr{R}_{abcd}$. Expanding the right side of (56) by the Leibniz rule,

$$\begin{aligned}\frac{1}{2}\mathscr{R}_{abcd}\, l^d = h_{[a}{}^r h_{b]}{}^s h_c{}^t [&(\nabla_r h_s{}^m) h_t{}^n \nabla_m l_n \\ &+ h_s{}^m (\nabla_r h_t{}^n) \nabla_m l_n + h_s{}^m h_t{}^n \nabla_r \nabla_m l_n]\end{aligned} \tag{57}$$

We consider the three terms on the right, one at a time. For the first term, $\nabla_r h_s{}^m = \nabla_r(\delta_s{}^m + \xi_s \xi^m) = \xi_s \nabla_r \xi^m + \xi^m \nabla_r \xi_s$. But ξ_s is annihilated by the $h_b{}^s$ in (57). Hence, this first term becomes $h_{[a}{}^r h_{b]}{}^s h_c{}^t [\xi^m (\nabla_r \xi_s) h_t{}^n \nabla_m l_n]$. But this vanishes, for $\nabla_r \xi_s$ is symmetric, while $h_{[a}{}^r h_{b]}{}^s$ is skew in "r" and "s". Thus, the first term on the right in (57) does not contribute. For the second term, we proceed similarly: $\nabla_r h_t{}^n = \xi_t \nabla_r \xi^n + \xi^n \nabla_r \xi_t$. The ξ_t is annihilated by $h_c{}^t$. Hence, using $\nabla_r \xi_t = \alpha h_{rt}$, this second term becomes $h_{[a}{}^r h_{b]}{}^s h_c{}^t [h_s{}^m \xi^n \alpha h_{rt} \nabla_m l_n]$. To simplify further, note that $\xi^n \nabla_m l_n = \nabla_m (\xi^n l_n) - l^n \nabla_m \xi_n = 0 - l^n \nabla_m \xi_n = -\alpha l_m$, where we have again used (46). Thus, using $h_a{}^m h_{mn} = h_{an}$, the second term on the right in (57) becomes, finally, $-\alpha^2 h_{c[a} l_{b]}$. Mercifully, the third term on the right in (57) is easy. It is just $h_{[a}{}^r h_{b]}{}^s h_c{}^t \nabla_r \nabla_s l_t$, which, by the definition of the Riemann tensor, is just $\frac{1}{2} h_a{}^r h_b{}^s h_c{}^t R_{rstd}\, l^d$. Substituting these evaluations in the right side of (57), we obtain

$$\frac{1}{2}\mathscr{R}_{abcd}\, l^d = \frac{1}{2} h_a{}^m h_b{}^n h_c{}^p R_{mnpd}\, l^d - \alpha^2 h_{c[a} h_{b]d} l^d \tag{58}$$

But, since l^d is an arbitrary spatial vector, (58) implies

$$\mathscr{R}_{abcd} = h_a{}^m h_b{}^n h_c{}^p h_d{}^q R_{mnpq} - 2\alpha^2 h_{c[a} h_{b]d} \tag{59}$$

Eqn. (59) relates the spatial curvature, $\mathscr{R}_{abcd}$, to the curvature of space-time, R_{abcd}, and α. The idea is to use (59) to determine the spatial curvature.

Contracting (59) over "b" and "d", we have

$$\begin{aligned}\mathscr{R}_{ac} &= h_a{}^m h_c{}^p h^{nq} R_{mnpq} - 2\alpha^2 h_{ac} \\ &= h_a{}^m h_c{}^p R_{mp} + h_a{}^m h_c{}^p \xi^n \xi^q R_{mnpq} - 2\alpha^2 h_{ac}\end{aligned} \tag{60}$$

Now we are in great shape. The Ricci tensor of space-time, R_{mn}, appears in (60), but we have an expression for it from Einstein's equation, (50). From (50), $h_a{}^m h_c{}^n R_{mn} = 4\pi G \rho\, h_{ac}$. The expression $R_{amcn}\, \xi^m \xi^n$ appears in (60), but we have essentially evaluated that in (48). Contracting (48) with ξ^b, we obtain $\mathscr{R}_{ambn}\, \xi^m \xi^n = -(\dot{\alpha} + \alpha^2) h_{ab}$. Thus, (60) becomes

$$\mathscr{R}_{ac} = h_{ac}[4\pi G \rho - \dot{\alpha} - 3\alpha^2] \tag{61}$$

Thus, the spatial Ricci tensor is a multiple of the spatial metric. Such a space (when the Ricci tensor is a multiple of the metric) is said to be of *constant curvature*.

To simplify (61) still further, substitute (53) for ρ, and $\dot{r}/r$ for α. Then, using $c = 2r^{-1} - (r)^2$, we have

$$\mathscr{R}_{ac} = 2r^{-2} c\, h_{ac} \tag{62}$$

The sign of the scalar curvature of space depends on the sign of c (the "binding energy"). For c positive (the universe recollapses), space is positively curved. It is natural that r^{-2} should appear on the right in (62). After all, r is a "characteristic size" of the universe, and $\mathscr{R}_{ac}$ is curvature, units $\sec^{-2}$. (If you blow up a balloon, the curvature of the balloon goes like r^{-2}, where r is the radius of the balloon.)

22. The Friedmann Solution: Continued

We begin with some remarks on the Friedmann solutions as a whole. We then discuss each of the three types individually.

The fundamental quantities in the Friedmann solutions are α (the expansion rate), ρ (the mass density of galaxies), and $\mathscr{R}$ ($\mathscr{R}^m{}_m$, the scalar curvature of space). These three quantities are not independent of each other, but satisfy a single identity;

$$\mathscr{R} = 16\pi\, G\, \rho - r\alpha^2 \tag{63}$$

(Proof: Contract (61), and use (51).) Thus, we may regard α and ρ as the independent quantities, with $\mathscr{R}$ expressed in terms of these by (63). If, at any point of space-time, the values of α and ρ are given, then the values of α and ρ for all t are thereby determined by (51) and (52). Thus, these two quantities would have to be measured at the present epoch of our universe to determine in which Friedmann solution we live, and in what portion of that solution we are.

The class of solutions obtained in Sect. (21) forms a one-parameter family. (It takes two measurements now to determine the structure of our Universe, but that gives us not only which solution we live in, but also where in the course of the evolution of that solution we are now.) To label this one-parameter family of solutions, note that Eqns. (51) and (52) admit a conserved quantity (i.e., a function of α and ρ which remains constant as α and ρ undergo (51) and (52).) This conserved quantity is

$$c = \left(\frac{4}{3}\pi\, G\, \rho\right)^{-2/3} \left[\frac{8}{3}\pi\, G\, \rho - \alpha^2\right] \tag{64}$$

(Proof: Eliminate r from $c = 2r^{-1} - (r)^2$.) As the solution evolves, the values of α and ρ change, but c remains a constant. Meanwhile, of course, $\mathscr{R}$ evolves via (63).

c negative. The universe expands from an initial singularity ($\rho = \infty$) (the "big bang"), and continues expanding forever thereafter ($\alpha > 0$). The rate of expansion decreases monotonically with time (physically, caused by the attraction of the galaxies for each other). In the early stages of the expansion (near the singularity) the "size of the universe", $r(t)$, goes like $r = \frac{1}{2}(6t)^{2/3}$. (This early behavior is not very significant physically. At this epoch, we expect the matter to be very dense, highly compressed. Thus, the matter should exert pressure at this epoch, so a fluid would be a better approximation to the matter then dust.) In the limit of large t, $r = (-c)^{1/2}t$. Thus, the rate of expansion approaches a constant at large t.

Space has constant negative curvature, i.e., we have hyperbolic space. The radius of curvature of space (r) decreases monotonically during the expansion.

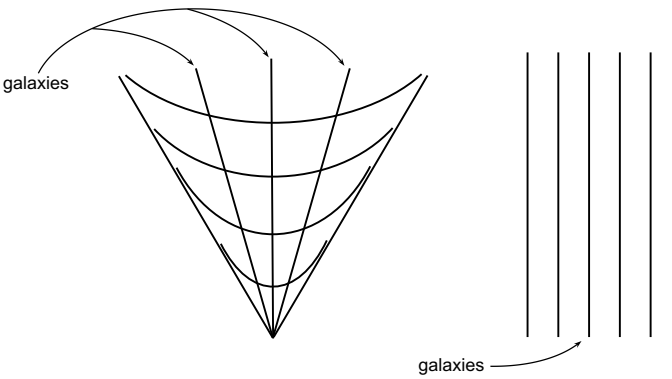
c zero. The universe expands from an initial singularity, and continues expanding thereafter, but at an ever decreasing rate. The size of the universe varies with t according to $r = \frac{1}{2}(6t)^{2/3}$. Thus, the rate of expansion is infinitive at the early epoch, and decrease to zero in the limit of large t.

Space has constant zero curvature. That is to say, space is flat.

c positive. The universe expands from an initial singularity, with $r = \frac{1}{2}(6t)^{2/3}$ at the earliest epochs. The rate of expansion is infinite in the earliest epochs. This

rate of expansion decrease, eventually passing through zero. Thus, the universe stops expanding, and begins contracting. The maximum occurs at time $t = \pi c^{-3/2}$ from the initial singularity. The universe recontracts, in the same amount of time, to a final singularity.

Space has constant positive curvature. Thus, the metric of space is that of 3-sphere. In particular, the universe is closed spatially. The radius of curvature of space is zero at the singularities, but increase to a maximum at the moment of maximum expansion. This maximum value is $\mathscr{R} = \frac{3}{2}c^3$.



Suppose we set $\rho = 0$ in the Friedmann solutions? Then (52) is satisfied identically, and (51) has solutions $\alpha = t^{-1}$ and $\alpha = 0$. In the first case, $\mathscr{R} = -6t^{-2}$; in the second, $\mathscr{R} = 0$. Thus, in the first case, we have an expanding Universe of negative spatial curvature; in the second, a static Universe of zero spatial curvature. It is not difficult to check that these are both Minkowski space, but sliced up in different ways.

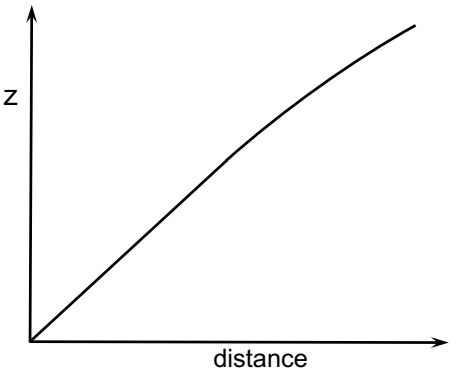
In the first case, the galaxies diverge from a common origin in Minkowski space. In the second case, the galaxies are parallel straight lines.

23. Friedmann Solutions: The Observations

We have seen that our universe appears to have the following properties: the dominant form of matter is dust (galaxies), and space is isotropic about each event in space-time. The only solutions of Einstin's equation which possess these features are the Friedmann solutions. Thus, we would expect that one of these Friedmann solutions describes the large- scale behavior of our own universe. Which one? What observations in our universe can be compared with the predictions of the Friedmann solutions? We shall be concerned with these questions in the present section.

We have seen in Sect. 22 that, in order to determine in which Friedmann solution we live, it is necessary to determine the values, at the present epoch, of ρ (mass density of the galaxies) and α (expansion rate). It is quite clear how (at least in principle) one would determine the present value of ρ. One would select a spatial volume (much larger then the inter-galactic distance, so local irregularities would be averaged out), determine the total mass inside that volume, and divide the mass by the volume. This is the present value of ρ. But what measurements would one make on our universe to determine the present value of α? We now begin the discussion of this – a much more complicated – question.

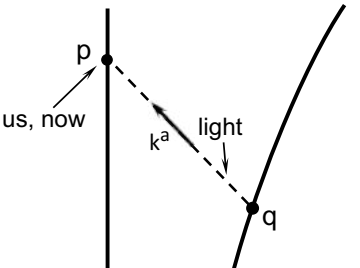
The following phenomenon is observed in our universe. The light reaching us from a distant galaxy is found to have a longer wavelength than the light had when it left the galaxy. (We know the wavelength the light had when it left the galaxy because it arose from atomic transitions.) The *red-shift* of a galaxy is defined by $z = (\lambda_r - \lambda_e)/\lambda_e$, where λ_e is the wavelength of the light emitted by the galaxy, and λ_r is the wavelength of the light we receive. Thus, $z = 0$ corresponds to no shifting of wavelengths. It is found, furthermore, that the red-shift of a galaxy depends on the distance of that galaxy from us. Thus, experimentally, one determines a graph such as the one below.



This graph is called the *observed distance-red-shift relation*.

The idea is now the following. Fix a Friedmann solution. Then, using this solution, we can compute a predicted distance-red-shift relation. By comparing the various predicted relations with the actual observed relation, one can make statements about which Friedmann solutions are active candidates to represent our own universe. It will turn out that this comparison yields only one useful piece of information: the value of α at the present epoch. Thus, our next task is to compute, within a given Friemann solution, what the distance-red-shift relation should be.

Fix a world-line of dust to represent "our galaxy". Fix a point p on this world-line to represent "us, now". Draw a null geodesic from p into the past to some point q on the world-line of another dust particle. This null



geodesic represents the light reaching us from the distant galaxy. The point q is the event "the galaxy emitted the light". Let k^a denote the tangent vector to this geodesic, so we have $k^a k_a = 0$ and $k^m \nabla_m k^a = 0$. This is the geometrical set-up. Two things must be discussed within this framework: the red-shifts, and the distances.

We begin with the red-shift. Let our light ray pass by a galaxy with four-velocity ξ^a. Then, as we saw in Sect.17, $\omega = k_a \xi^a$ is the (angular) frequency of the light as seen by an observer on the galaxy. In particular, ω, evaluated at q, is the emitted frequency of the light, while ω, evaluated at p, is the received frequency of the light. Let a prime denote $k^m \nabla_m$ (i.e., the derivative along the light ray). Then

$$\begin{aligned}\omega' &= k^m \nabla_m \omega = -k^m \nabla_m (\xi^a k_a) = -\xi^a k^m \nabla_m k_a - k_a k^m \nabla_m \xi_a \\ &= 0 - \alpha k^a k^b (g_{ab} + \xi_a \xi_b) = -\alpha k^a k_a - \alpha (k^a \xi_a)^2 = -\alpha \omega^2\end{aligned}$$

where we have used (46). The meaning of this equation is clear physically. If α is positive, then the frequency of the light (as seen by intervening galaxies) decrease on moving from q to p. That is, the observed wavelength will be larger than the emitted wavelength.

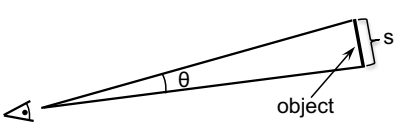
Thus, if we have our Friedmann solution (so we know α everywhere), and if we know the frequency of the light emitted at q, then we can determine ω along our entire light ray from

$$\omega' = -\alpha \omega^2 \tag{65}$$

In particular, we can determine the value of ω at p, i.e., the frequency of the light received. These frequencies determine the corresponding wavelengths. Hence, from (65) we can determine the red-shift z for the other galaxy.

We now have half of what we need to draw the theoretical distance-red-shift relation. We know how to compute the redshift of the galaxy through event q. We next must determine the distance of that galaxy.

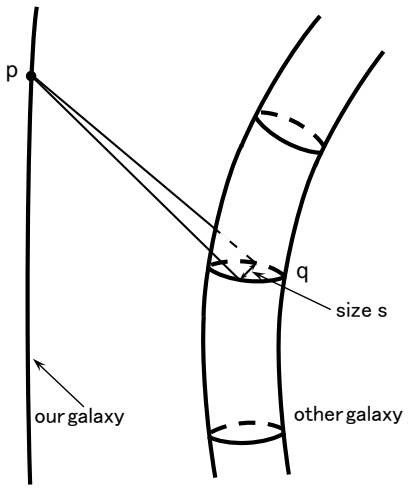
Now distance is not a "natural" concept in general relativity. That is, there is no simple, unambiguous quantity associated with the points p and q in the earlier figure which everyone would agree to call the "spatial distance" between the galaxies. This type of situation arises frequently in general relativity. What we must do is analyze in detail the actual measurements and computations the astronomer makes in arriving in the quantity he chooses to call the "distance" between the galaxies. We then express what the astronomer actually did in terms of tensors in our Friedmann solution. Thus, we acquire the ability to compute, for our given Friedmann solution, a quantity (associated with p and q) which can be called the distance between the galaxies. In practice, what the astronomer does to determine what he calls distance is rather complicated. It is convenient to over-simplify. We select one reasonable operational definition of distance, and analyze it within the Friedmann solutions.



In ordinary Euclidean space, if one looks at an object of actual size s, and if it subtends an angle θ (small), then we would say we are a distance s/θ from the object. One might (and often does) use a similar technique in astronomy. One looks at a galaxy, knowing (say, from the study of galactic structure) its size s. One measures the apparent

angular size of the galaxy, and thus computes its "distance". We shall adopt this as our definition of distance.

Thus, from the point "p" of our Friedmann solution, we draw two nearby null geodesics. These reach opposite sides of the galaxy located at q. Let s be the actual diameter of this galaxy. Our task is to compute θ, the angle we see these light rays reaching us (at p) with. Then s/θ will be the "distance" of this galaxy.



We are dealing with two nearby null geodesics. Clearly, therefore, the way to compute θ from s is to use the equation of geodesic deviation.

Let l^a be a vector, defined at each point of our null geodesic, which is parallel transported along the geodesic. That is, we have $k^m \nabla_m l_a = 0$ Hence,

$$\begin{aligned} k^m \nabla_m (l^a l_a) &= 2 l^a k^m \nabla_m l_a = 0 \\ k^m \nabla_m (l^a k_a) &= k_a (k^m \nabla_m l^a) + l^a (k^m \nabla_m k_a) = 0 \\ k^m \nabla_m (l^a \xi_a) &= \xi_a k^m \nabla_m l^a + l^a k^m \nabla_m \xi_a = 0 + l^a k^m \alpha (g_{am} + \xi_a \xi_m) \\ &= \alpha (l^m k_m) + \alpha (\xi_a l^a)(\xi_m k^m) \end{aligned}$$

Clearly, we can choose $l^a l_a = 1$, $k^a l_a = 0$ and $\xi^a l_a = 0$. We do so. We now ask that $\kappa\, l_a$ satisfy the equation of geodesic deviation, where κ is a function along the geodesic. That is, we ask that

$$k^m k^n \nabla_m \nabla_n (\kappa l_a) = -\kappa R_{manb}\, k^m k^n l^b \tag{66}$$

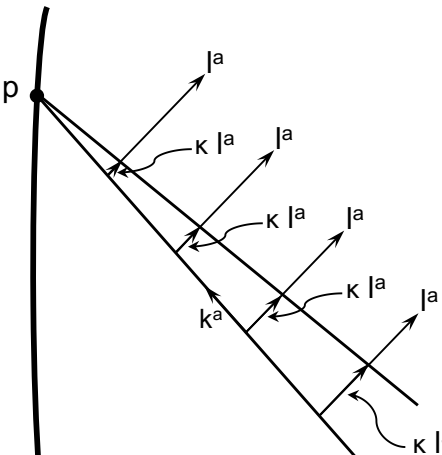
Using $k^m \nabla_m l_a = 0$, the left side becomes $\kappa'' l_a$. To evaluate the right side, note that (50) together with isotropy implies

$$R_{abcd} = \frac{16}{7} \pi\, G\, \rho [g_{a[c} g_{d]b} - 3 \xi_{[a} g_{b][c} \xi_{d]}] \tag{67}$$

Hence, the right side of (66) is $-4\pi\, G\, \rho\, \omega^2 \kappa\, l_a$. Getting rid of the l_a's, Eqn. (66) becomes

$$\kappa'' = -4\pi\, G\, \rho\, \omega^2 \kappa \tag{68}$$

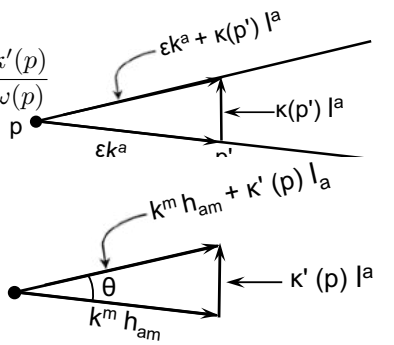
Geometrically, the situation is now the following. The vector l^a is just a unit direction. But $\kappa\, l^a$ satisfied the equation of geodesic deviation provided κ varies along our geodesic according to (68). Hence, κ represents the distance between the two nearby geodesics. We are concerned with the solution κ of (68) which satisfies the two boundary conditions: $\kappa(q) = s$ (the size of the other galaxy), and $\kappa = 0$ (the two nearby light rays meet



at p). These two boundary conditions determine a unique solution of (68).

We now have a description of the two nearby light rays. Only one task remains; to determine the apparent angular size θ of the galaxy as seen from p. To do this, we first find the tangent vectors to the two null geodesics, at p. Move a parameter distance ϵ ($\ll 1$) along each geodesic from p. For the first geodesic, this infinitesimal displacement is represented by $\epsilon\, k^a$. For the second geodesic, the displacement is represented by $\epsilon\, k^a + \kappa(p')\, l^a$ where p' is the point displaced $\epsilon\, k^a$ from p. (This is the definition of geodesic deviation.) But $\kappa(p) = 0$. Hence, $\kappa(p') = \epsilon\, k^m \nabla_m \kappa$ (at p) $= \epsilon\, \kappa'(p)$. Hence, the tangent vectors to the two light rays are k^a and $k^a + \kappa'(p)\, l_a$. To obtain the apparent spatial direction from which we (at p) see these light rays coming in, we project orthogonal to ξ^a (the four-velocity of our galaxy). Hence, the light rays appear to come in in spatial directions $h_a{}^m k_m$ and $h_a{}^m(k_m + \kappa'(p) l_m = h_a{}^m k_m + \kappa'(p) l_a$. What is the angle between these two nearby spatial vectors? From Euclidean geometry, it is

$$\theta = [h^{am} k_m h_a{}^n k_n]^{-1/2} [\kappa'(p) l_a \kappa'(p) l^a]^{1/2} = \frac{\kappa'(p)}{\omega(p)} \tag{69}$$



We are now prepared to summarize what is a rather complicated situation. We first pick a number s to represent the size of the galaxy at q. We then solve (68) for κ along our null geodesic, subject to $\kappa(p) = 0$ and $\kappa(q) = s$. Substituting κ' into (69), we obtain the angle subtended by the galaxy. Then s/θ is what we call the distance of that galaxy. (Note that this distance is independent of s. Since (68) is linear, the function κ defined by (68) is proportional to s. Hence, from (69), θ is proportional to s. So, s/θ is independent of s. This, of course, is what we would expect.)

We now have a prescription for computing the theoretical distance-red-shift relation. Choose a point p (to represent us, now), and a null geodesic into the past from p to some point q. Solve (65) to determine the redshift of the light from this galaxy. Then pick s (it makes no difference what s), and solve (68) subject to $\kappa(p) = 0$ and $\kappa(q) = s$. Find θ from (69), and set distance $= s/\theta$. Thus, once we pick q, we can compute the redshift z and the distance s/θ. Repeating for various choices of q (the event of light leaving the other galaxy), we obtain the theoretical distance-red-shift relation.

In fact, we shall not attempt to solve these equations. (The solutions are complicated, not very illuminating, and not very useful in practice.) Instead, we ask a simpler and more important question. What is the slope of the distance-red-shift curve at the origin (i.e., the value of dz/d(distance) at $z = 0$? The observed distance-red-shift curve is nearly a straight line, so the only quantity which has been determined observationally with any accuracy is this slope. Let us compute the slope of the theoretical distance-red-shift curve.

Let q lie a parameter-amount ϵ ($\epsilon \ll 1$) along the null geodesic from p, so the connecting vector is $\epsilon\, k^a$. Then since $\kappa(p) = 0$, $\kappa(q) = \epsilon\, \kappa'(p)$, which is that we call s. Hence, from (69), $\theta = \kappa'(p)/\omega(p) = (s/\epsilon)/\omega(p)$. So, distance $s/\theta = \epsilon\, \omega(p)$. (Note that, to first order, we don't even need to use (68).) Thus, the distance between the two galaxies in this case is $\epsilon\omega(p)$. We now want to compute the redshift between these two galaxies. From (65), $\omega(p) = \omega(q) - \epsilon\, \alpha(p)\, \omega^2(p)$, to first order in ϵ. Hence,

$$z = \frac{\lambda_{rec} - \lambda_{emit}}{\lambda_{emit}} = \frac{1/\omega(p) - 1/\omega(q)}{1/\omega(q)} = \frac{\omega(q)}{\omega(p)} - 1$$
$$= \frac{\omega(p) + \epsilon\,\alpha(p)\,\omega^2(p)}{\omega(p)} - 1 = \epsilon\,\alpha(p)\,\omega(p)$$

to first order in ϵ. Taking the ratio, $z/distance = \alpha(p)$. Thus, *the slope of the theoretical distance-red-shift curve at the origin is precisely the value of* α *at the observation point* p. (It's clear, from the definition of α and the remark surrounding it, that this is the only reasonable answer. This is why we called α the "rate of expansion".)

Thus, the value of α in our universe at the present epoch is determined merely by measuring the slope of the distance-red-shift curve (observational) at the origin.

We can now determine the two quantities – ρ and α at the present epoch – necessary to determine the Friedmann solution in which we live. What do the observations say? Firstly, this α – called the *Hubble constant* – is known relatively well. The best current value is $\alpha = 1.7 \times 10^{-19}\,\mathrm{sec}^{-1}$ (i.e., the universe expands one part in 2×10^{11} per year). That's good perhaps to a factor of five. The current value of ρ is known much less accurately. What is the most interesting question? Clearly, it's whether the constant c in (63) is positive, negative, or zero (i.e., whether our universe is closed spatially and will later contract, or whether it is open and will continue expanding). Set

$$\rho_c = \frac{\alpha^2}{\frac{8}{3}\pi\,G} = 170\ \mathrm{g/sec}^3 \quad (= 6.3 \times 10^{-30}\ \mathrm{g/cm}^3) \tag{70}$$

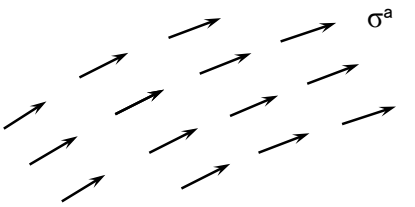
the critical density. Then, if ρ (observed) $> \rho_c$, we have the spatially closed universe, while if ρ (observed) $< \rho_c$, we have the open universe. The best current value for ρ is $\rho = (1/30)\rho_c$. This value suggests the spatially open, ever expanding universe. Unfortunately, the observational error in this ρ is such that no definite conclusion can be drawn, at the present time, on this question.

It's all very unsatisfactory. General relativity makes very definite, detailed predictions. Any reasonable theory would give predictions of the same general order of magnitude. The observations, as a whole, are good to factors such as ten. We cannot even make a definite statement, from the observations, as to whether our universe should be spatially open or closed.

24. Symmetries

The notion of a symmetry – a motion in space-time under which all the physics is invariant – is an important one. It is often the case in practice that essentially all the simple and useful information about a space-time arises from the presence of symmetries.

Recall the motion of the Lie derivative. We have a vector field σ^a on a manifold M. Suppose we displace each point of M infinitesimally by an amount determined by the value of σ^a at that point. Thus, σ^a defines a "motion", an infinitesimal "sliding along" in the direction of σ^a. Now let $T^{a\cdots c}{}_{b\ldots d}$ be any tensor field in M. Then one obtain two tensors fields: the original $T^{a\cdots c}{}_{b\ldots d}$ and the field which results from subjecting $T^{a\cdots c}{}_{b\ldots d}$ to this "sliding along" by σ^a. The second field minus the first, divided by the (parameter) amount of sliding along, gives the *Lie derivative* of $T^{a\cdots c}{}_{b\ldots d}$. It follows that the Lie derivative of $T^{a\cdots c}{}_{b\ldots d}$ by σ^a is given by the right side of



$$\begin{aligned}\pounds_\sigma T^{a\ldots c}{}_{b\ldots d} &= \sigma^m\nabla_m T^{a\ldots c}{}_{b\ldots d} - T^{m\ldots c}{}_{b\ldots d}\,\nabla_m\sigma^a - \cdots \\ &- T^{a\ldots m}{}_{b\ldots d}\,\nabla_m\sigma^c + T^{a\ldots c}{}_{m\ldots d}\,\nabla_b\sigma^m + \cdots + T^{a\ldots c}{}_{b\ldots m}\,\nabla_d\sigma^m \end{aligned} \tag{71}$$

A vector field σ^a in space-time is said to define a *symmetry* if all the tensor fields of physical interest on the space-time (i.e., the metric, and the fields which describe the various types of matter in the space-time) are invariant under the "sliding" by σ^a. That is, σ^a defines a symmetry if $\pounds_\sigma$ (all physical fields)$= 0$.

In particular, if σ^a is a symmetry, then $\pounds_\sigma g_{ab} = \sigma^m\nabla_m g_{ab} + g_{mb} + g_{am}\sigma^m = 2\nabla_{(a}\sigma_{b)}$. This equation,

$$\nabla_{(a}\sigma_{b)} = 0 \tag{72}$$

is called *Killing's equation*, the solutions *Killing vectors*. (Thus, a symmetry vector field is always a Killing vector field. It usually turns out that Killing vectors are also symmetries.)

Note, that, since (72) is linear, any linear combination of Killing vectors (with constant coefficients) is again a Killing vector. Another elementary consequence of Killing's equation is the following . Let σ^a be a Killing vector, and set $\kappa = \sigma^b\sigma_b$. Then

$$\begin{aligned}\sigma^a\nabla_a\kappa &= -\sigma^a\nabla_a(\sigma^b\sigma_b) = 2\sigma^a\sigma^b\nabla_a\sigma_b \\ &= 2\sigma^a\sigma^b\nabla_{(a}\sigma_{b)} = 0\end{aligned}$$

where we have used (72) in the last step. That is, the norm of a Killing vector is constant along the integral curves of σ^a. Furthermore, if σ^a is a Killing vector, then $\sigma^m\nabla_m\sigma_a = -\sigma^m\nabla_a\sigma_m = -\frac{1}{2}\nabla_a\kappa$, where we have used (72) in the first step. Thus, the curvature of an integral of a Killing vector is $-\frac{1}{2}$ times the gradient of the norm κ. (Therefore: Lemma: The integral curves of a Killing vector are geodesics when and only when the Killing vector has constant norm.)

We now derive what is perhaps the most important single consequence of (72). Let σ^a be a Killing vector. Then

$$\begin{aligned}
&\nabla_a\nabla_b\,\sigma_c = -\nabla_a\nabla_c\,\sigma_b = -\nabla_c\nabla_a\,\sigma_b - R_{acb}{}^m\,\sigma_m \\
&= \nabla_c\nabla_b\,\sigma_a - R_{acb}{}^m\,\sigma_m = \nabla_b\nabla_c\,\sigma_a + R_{cba}{}^m\,\sigma_m - R_{acb}{}^m\,\sigma_m \\
&= -\nabla_b\nabla_a\,\sigma_c + R_{cba}{}^m\,\sigma_m - R_{acb}{}^m\,\sigma_m \\
&= -\nabla_a\nabla_b\,\sigma_c - R_{bac}{}^m\,\sigma_m + R_{cba}{}^m\,\sigma_m - R_{acb}{}^m\,\sigma_m
\end{aligned}$$

where, in the first, third and fifth steps, we have used (72), and in the second, fourth, and sixth steps we have used the definition of the Riemann tensor. Adding $\nabla_a\nabla_b\,\sigma_c$ to both sides (i.e., to the first and last expression) of this equation, and using $R_{acb}{}^m + R_{cba}{}^m + R_{bac}{}^m = 0$ (a Riemann tensor symmetry) we obtain $2\nabla_a\nabla_b\,\sigma_c = 2R_{cba}{}^m\,\sigma_m$. But $R_{cba}{}^m = R^m{}_{abc}$ (Riemann tensor symmetry). Hence, we obtain

$$\nabla_a\nabla_b\,\sigma_c = R^m{}_{abc}\,\sigma_m \tag{73}$$

This is the desired result. (Note that the result of antisymmetrizing (73) over "a" and "b" is an identity.) That is, if the first derivative of a vector field is restricted by (72), then its second derivative can be expressed (via (73)) in terms of the original vector field. We need never deal directly with second or higher derivatives of a Killing vector.

We now derive two consequences of (73), Let σ^a and γ^a be Killing vectors, and set

$$\rho^a = [\sigma,\gamma]^a = \sigma^m\nabla_m\gamma^a - \gamma^m\nabla_m\sigma^a \tag{74}$$

We show that ρ^a is also a Killing vector. Taking a derivative of ρ^b:

$$\begin{aligned}
&\nabla^a\rho^b = (\nabla^a\sigma^m)(\nabla_m\gamma^b) + \sigma^m\nabla^a\nabla_m\gamma^b - (\nabla^a\gamma^m)(\nabla_m\sigma^b) - \gamma^m\nabla^a\nabla_m\sigma^b \\
&= [(\nabla^a\sigma^m)(\nabla_m\gamma^b) - (\nabla^a\gamma^m)(\nabla_m\sigma^b)] + [\sigma^m\nabla^a\nabla_m\gamma^b - \gamma^m\nabla^a\nabla_m\sigma^b] \\
&= [-(\nabla^m\sigma^a)(\nabla_m\gamma^b) + (\nabla^m\gamma^a)(\nabla_m\sigma^b)] + [\sigma^m R^{na}{}_{m}{}^{b}\,\gamma_n - \gamma^m R^{na}{}_{m}{}^{b}\,\sigma_n] \\
&= [-(\nabla^m\sigma^a)(\nabla_m\gamma^b) + (\nabla^m\sigma^b)(\nabla_m\gamma^a)] + [\sigma^m\gamma^n R_n{}^a{}_m{}^b - \sigma^m\gamma^n R_m{}^b{}_n{}^a\,]
\end{aligned}$$

where, in the first step, we have substituted (74); in the second, we have rearranged terms; in the third, we have used (72) in the first bracket and (73) in the second; and in the fourth we have reversed the roles of (the dummy indices) "m" and "n" and used $R_{abcd} = R_{cdab}$ in the last term in the second bracket. Now consider the right side of this expression. Each bracket is, evidently, antisymmetric in "a" and "b". Hence, $\nabla^{(a}\rho^{b)} = 0$, so ρ^a is a Killing vector.

The vector field $[\sigma,\gamma]^a$ is called the *Lie Bracket*, or *commutator*, of σ^a and γ^a. Evidently, $[\sigma,\gamma]^a = -[\gamma,\sigma]^a$. Furthermore , the Lie bracket is linear (using linear combinations with constant coefficients) in σ^a and γ^a. Finally, it is not difficult to check that

$$[[\sigma,\gamma],\lambda]^a + [[\gamma,\lambda],\sigma]^a + [[\lambda,\sigma],\gamma]^a = 0$$

(The remarks to here in this paragraph are true for any vector fields – whether or not they are Killing vectors.) Thus, the Killing vectors on a space-time have the structure of a Lie algebra (a vector space with an antisymmetric bracket operation, subject to the Jacobi identity).

As a second example of the use of (73), we find all Killing vectors in flat space (of any signature and dimension). Let σ^a be a Killing vector. Then, since $R_{abcd} = 0$ (flat space), (73) gives $\nabla_b\sigma_c = 0$. Hence,

$$\nabla_a \sigma_b = \underline{F}_{ab} \tag{75}$$

where $\underline{F}_{ab}$ is a constant tensor field. Eqn. (72) implies that $\underline{F}_{ab}$ is antisymmetric. To integrate (75), introduce an origin O in our flat space, and let x^a be the vector field whose value at each point is the position vector of that point relative to O. (Precisely, x^a is defined by the properties that x^a vanishes at O, and $\nabla_a x^b = \delta_a{}^b$.) The general solution of (75) is thus

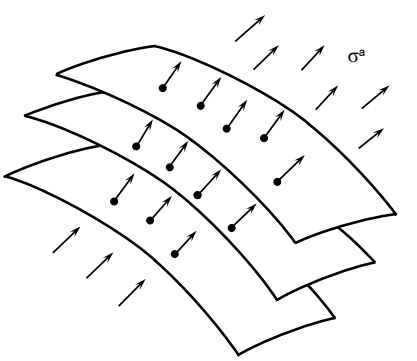
$$\sigma_b = x^a \underline{F}_{ab} + \underline{\sigma}_b \tag{76}$$

where $\underline{\sigma}_b$ is a constant vector field. Thus, the general Killing vector in flat space is defined by a constant antisymmetric tensor field $\underline{F}_{ab}$ and a constant vector field $\underline{\sigma}_b$. (The particular constant fields which represent a given Killing vector depend on the choice of origin.)

The Killing vectors $\sigma^a = \underline{\sigma}^a$ are called translations. If our flat space is Minkowski space, a translation is said to be temporal or spatial according as $\underline{\sigma}^a$ is timelike or spacelike. (Of course, $\underline{\sigma}^a$ could also be null in Minkowski space.) Consider Eucliden 3-space. Then $\underline{F}_{ab} = \epsilon_{abc} W^c$, for some constant vector field W^c. The corresponding Killing vector, $\sigma_b = \underline{F}_{ab} x^a$, is called a rotation about the origin O, with axis of rotation W^a. Consider Minkowski space. Then $\sigma_b = \underline{F}_{ab} x^a$ is said to generate a Lorentz transformation about O. To decompose Lorentz transformations, introduce a unit timelike vector t^a at O. Then, since $\underline{F}_{ab}$ is antisymmetric, we can introduce the "electric" and "magnetic" parts of $\underline{F}_{ab}$, with respect to t^a. (The equations are precisely (20), (21), and (22)). The electric part of $\underline{F}_{ab}$, $t^b \underline{F}_{ab}$ is said to generate a boost, or velocity transformation, in the direction $t^b \underline{F}_{ab}$. The magnetic part is said to generate a spatial rotation about the axis $\frac{1}{2}\epsilon_{abcd} t^b \underline{F}^{cd}$. Thus, the decomposition of Lorentz transformations into boosts and rotations depends on the choice of a time-direction t^a.

A vector field σ_a (not necessarily a Killing vector) in space-time is said to be *hypersurface-orthogonal* if

$$\sigma_{[a} \nabla_b \sigma_{c]} = 0 \tag{77}$$



Geometrically, (77) is the condition that there exist a family of three-dimensional surfaces in space-time to which σ_a is everywhere orthogonal. (The fibers of a twisted rope define a vector fields in Euclidean 3-space which is not hypersurface-orthogonal. Translations are always hypersurface-orthogonal.)

We let σ^a be a hypersurface-orthogonal Killing vector, i.e., let σ^a satisfy (72) and (77). (About half of all Killing vectors one meets in practice are hypersurface-orthogonal.) Contracting (77) with σ^a, using $\sigma^a \nabla_a \sigma_b = -\sigma^a \nabla_b \sigma_a = \frac{1}{2} \nabla_a \kappa$, where, $\kappa = \sigma^c \sigma_c$, we obtain (when $\kappa \neq 0$)

$$\nabla_a \sigma_b = -\kappa^{-1} \sigma_{[a} \nabla_{b]} \kappa \tag{78}$$

Thus, the value of the derivative of a hypersurface-orthogonal Killing vector at a point is expressible in terms of the values of the Killing vector, its norm, and the gradient of its norm at the point. This equations, and the result of substituting it

into (73), are frequently the keys to unraveling a space-time with a hypersurface-orthogonal Killing vector.

All we have done so far in this section is introduce some definitions and some useful mathematical techniques. What is the physics of Killing vectors? We have seen that Killing vectors are associated with symmetries, and of course symmetries are usually associated with conserved quantities. So Killing vectors should lead to conserved quantities. These come in two varieties.

Let σ^a be a Killing vector in space-time, and consider a geodesic with tangent vector p^a. Then

$$\begin{aligned} p^m \nabla_m (p^a \sigma_a) &= \sigma_a p^m \nabla_m p^a + p^a p^m \nabla_m \sigma_a \\ &= 0 + p^a p^m \nabla_{(m} \sigma_{s)} = 0 \end{aligned}$$

Hence, $p_a \sigma^a$ is constant along geodesic. That is, $p_a \sigma^a$ is a "constant of the motion of the geodesic". (Not a very good term, since geodesics do not "move". (Nothing "moves" in space-time.)) (It is easily checked that if $p_a \sigma^a$ is constant along the geodesic, for every geodesic, then σ^a is a Killing vector.)

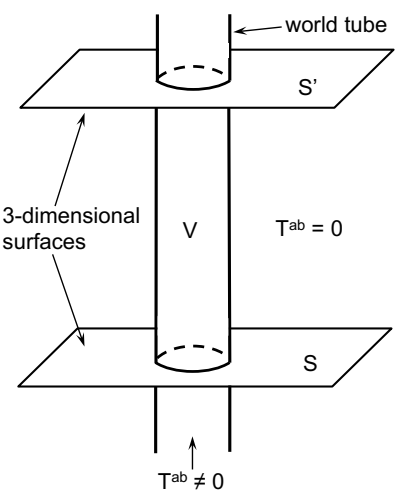
The second type of conserved quantity arises from the stress-energy T^{ab}. Let σ^a be a Killing vector. Then

$$\nabla_a (T^{ab} \sigma_b) = \sigma_b \nabla_a T^{ab} + T^{ab} \nabla_a \sigma_b = 0$$

That is, $T^{ab}\sigma_b$ is divergence-free. Now suppose we have a star, or some isolated body. We integrate $T^{ab}\sigma_b$ over a three-dimensional surface S cutting the world-tube of the body. If we were to integrate instead over some other surface S', then the difference between the integrals would be

$$\int\limits_{S'} T^{ab}\, \mathrm{d}S_a - \int\limits_{S} T^{ab}\sigma_b\, \mathrm{d}S_a = \int\limits_{V} \nabla_a (T^{ab}\sigma_b)\, \mathrm{d}V = 0$$

where V is the portion of the world-tube between S and S', Hence, $\int\limits_S T^{ab}\sigma_a\, \mathrm{d}S_a$ is just a number, independent of the surface S. It is a conserved quantity ("conserved" in the sense that the value of the integral is independent of the surface).



In a general space-time, there will be no Killing vectors (as we would expect: most systems have no symmetries). Killing vectors are important, not because they arise commonly in actual systems, but because permit us to do something (e.g., with Einsten's equation) where otherwise we might be able to do nothing. The neat decomposition of Killing vectors in Minkowski space into translations, rotations, and boosts is not available for Killing vectors in a curved space-time. A Killing vector is just a Killing vector. However, often a Killing vector in a curved space-time closely resembles, in certain respects, a Killing vector in Minkowski space. (This is usually the case when the space-time is asymptotically flat, so one can examine the asymptotic behavior of the Killing vector.) When such a resemblances exists, it is usually emphasized by calling the Killing vector by the name of its Minkowski space analog (e.g., calling it a temporal translation, a rotation, etc.).

If a Killing vector in space-time is called a temporal translation, then the conserved quantity $p_a\sigma^a$ (where p_a is the four-momentum of a particle on the geodesic) is called the energy of the particle, while the conserved quantity associated with T^{ab} is called the energy of the star. Similarly, the conserved quantities associated with Killing vectors called spatial translations are called components of the momentum, and the conserved quantities associated with Killing vectors called spatial rotations are called components of the angular momentum. (For some reason, the conserved quantities associated with boosts do not seem to be very useful. They have no name.) Examples: Consider Minkowski space, with all its Killing vectors. Then what we above called the energy, momentum, and angular momentum associated with a p^a or T^{ab} are precisely what is called energy, momentum, and angular momentum in special relativity. (Note that these concepts from special relativity are, in a sense, simpler when introduced from the point of view of general relativity. It is clear why conserved quantities arise (because of symmetry). Furthermore, we don't have to prove that things "transform like a 4-vectors".)

Certain intuitive "symmetry expressions" are given a precise meaning in general relativity. We introduce four such. A space-time is said to be "stationary" if it possesses a Killing vector which is everywhere timelike. A space-time is said to be "static" if it possesses a Killing vector which is hypersurface-orthogonal and everywhere timelike. (Physically, a stationary system presents the same aspect at each time, although there may be motion (e.g., water flowing through a pipe, or a rotating disc), while in a static system nothing moves.) A space-time is said to be "axially symmetric" if it possesses a Killing vector which is everywhere spacelike, and whose integral curves are closed curves. (The axis is the points of space-time where this Killing vector vanishes. It is normally a two-surface – but it is possible that there exist no such points!) Finally, a space-time is said to be "spherically symmetric" if it possess three Killing vectors, $\underset{1}{l}{}^a$, $\underset{2}{l}{}^a$, and $\underset{3}{l}{}^a$ (physically, which generate rotations about the three axes), all of which are everywhere spacelike, which are linearly dependent at each point (i.e., $\underset{1}{l}{}^{[a}\underset{2}{l}{}^b\underset{3}{l}{}^{c]} = 0$) , and which have the same commutation relations as do the generators of the rotation group:

$$[\underset{1}{l}, \underset{2}{l}]^a = \underset{3}{l}{}^a, \quad [\underset{2}{l}, \underset{3}{l}]^a = \underset{1}{l}{}^a, \quad [\underset{3}{l}, \underset{1}{l}]^a = \underset{2}{l}{}^a \tag{79}$$

(It is easily checked that the three Killing vectors which generate rotations about an origin in Euclidean three-space have these properties.)

There is a patters to these definitions. In each case, a term means the existence of Killing vectors which display as many features reminiscent of that term as are simple and meaningful in curved space-time.

25. The Schwarzschild Solution

Physically, the Schwarzschild solution represents the geometry of an "isolated, non rotating star, which has settled down to equilibrium". What properties would we expect such a solution to have? Firstly, we would expect the solution to be static, i.e., we would expect to have a timelike, hypersurface-orthogonal Killing vector t^a. Secondly, we would expect the solution to be spherically symmetric, i.e., we would expect to have Killing vectors $\underset{1}{l}{}^a$, $\underset{2}{l}{}^a$, and $\underset{3}{l}{}^a$ which are spacelike, linearly dependent, and have the commutation relations (79). Finally, we would expect that the time-translations and rotations commute, i.e., we would expect to have additional commutation relations

$$[t, \underset{1}{l}]^a = [t, \underset{2}{l}{}^a] = [t, \underset{3}{l}{}^a] = 0 \tag{80}$$

To summarize, we are concerned with space-time having four Killing vectors, with the commutation relations (79) and (80). For the matter composing the star, we take a fluid. Thus, we have the density ρ, pressure p, and (unit) velocity field η^a. Since the star is supposed to have "settled down to equilibrium", we suppose that the fluid does not "move relative to static observers", i,e., we take η^a a multiple of t^a.

To summarize, the Schwarzschild solution is a space-time with four Killing vectors, t^a (timelike, hypersurface-orthogonal), and $\underset{1}{l}{}^a$, $\underset{2}{l}{}^a$, $\underset{3}{l}{}^a$ (spacelike, linearly dependent), subject to (79) and (80), where the matter is a fluid with four-velocity field proportional to t^a. We now discuss the geometry of the Schwarzschild solution.

Set $\lambda = t^a t_a$, (so $\lambda < 0$ since t^a is timelike). We begin with the physical interpretation of this λ. Since t^a is hypersurface-orthogonal, we have, from (78)

$$\nabla_b t_c = -\lambda^{-1} t_{[b} \nabla_{c]} \lambda \tag{81}$$

Hence,

$$\begin{aligned} R^m{}_{abc}\, t_m &= \nabla_a \nabla_b\, t_c = \nabla_a(-\lambda^{-1}\, t_{[b} \nabla_{c]} \lambda) \\ &= \lambda^{-2} (\nabla_a \lambda)\, t_{[b} \nabla_{c]} \lambda - \lambda^{-1} (\nabla_a\, t_{[b}) \nabla_{c]} \lambda - \lambda^{-1} (\nabla_a \nabla_{[c} \lambda) t_{b]} \\ &= \frac{1}{2} \lambda^{-2} \nabla_a \lambda\, t_{[b} \nabla_{c]} \lambda - \lambda^{-1} (\nabla_a \nabla_{[c} \lambda) t_{b]} \end{aligned}$$

where, in the first step, we have used (73); in the second, we have substituted (81); in the third, we have expanded using the Leibniz rule, and in the fourth, we have again used (81).

Contracting the previous equation over "a" and "c",

$$\begin{aligned} R^m{}_b\, t_m &= \frac{1}{2} \lambda^{-2} (\nabla^c \lambda) t_{[b} \nabla_{c]} \lambda - \lambda^{-1} (\nabla^c \nabla_{[c} \lambda) t_{b]} \\ &= \frac{1}{4} \lambda^{-2} (\nabla^c \lambda)(t_b \nabla_c \lambda - t_c \nabla_b \lambda) - \frac{1}{2} \lambda^{-1} ((\nabla^c \nabla_c \lambda) t_b - (\nabla^c \nabla_b \lambda) t_c) \\ &= \frac{1}{4} \lambda^{-2}\, t_b (\nabla^c \lambda \nabla_c \lambda) - \frac{1}{2} \lambda^{-1}\, t_b \nabla^2 \lambda + \frac{1}{2} \lambda^{-1} t^c \nabla_c \nabla_b \lambda \end{aligned} \tag{82}$$

where, in the second step, we have expanded the antisymmetrizations, and in the third step, we have used $t^c \nabla_c \lambda = 0$ and written ∇^2 for $\nabla^c \nabla_c$. The last term on the right in (82) can be simplified further:

$$t^c\nabla_c\nabla_b\lambda = t^c\nabla_b\nabla_c\lambda = \nabla_b(t^c\nabla_c\lambda) - (\nabla_c\lambda)\nabla_b t^c$$
$$= \nabla_b(0) - \nabla^c\lambda\nabla_b t_c = -\nabla^c\lambda(-\lambda^{-1}t_{[b}\nabla_{c]}\lambda)$$
$$= \frac{1}{2}\lambda^{-1}t_b(\nabla^c\lambda\nabla_c\lambda)$$

Hence, (82) becomes

$$R_{mb}\,t^m = \frac{1}{2}\lambda^{-2}t_b(\nabla^c\lambda\nabla_c\lambda) - \frac{1}{2}\lambda^{-1}t_b\nabla^2\lambda \tag{83}$$

Eqn. (83) gives components of the Ricci tensor in terms of λ. We proceed (as always in such situations) to eliminate the Ricci tensor in favor of the matter variables. For a perfect fluid, Einstein's equation becomes

$$R_{ab} = 8\pi\,G[-\lambda^{-1}(\rho+p)t_at_b + \frac{1}{2}(\rho-p)g_{ab}] \tag{84}$$

where we have used the fact that the velocity field of the fluid is $(-\lambda)^{-1/2}t^a$ (i.e., that multiple of t^a which is unit). From (84), $R_{ab}\,t^b = -4\pi\,G\,t_a(\rho+3p)$. Substituting, we have from (83)

$$\lambda^{-1}\nabla^2\lambda - \lambda^{-2}(\nabla^2\lambda\nabla_c\lambda) = 8\pi\,G(\rho+3p) \tag{85}$$

This can be rewritten in the more suggestive form

$$\nabla^2[\frac{1}{2}\ln(-\lambda)] = 4\pi\,G(\rho+3p) \tag{86}$$

Eqn. (86) has a simple physical interpretation. We regard $\frac{1}{2}\ln(-\lambda)$ as analogous to the "Newtonian potential " φ, so (86) in analogous to the Newtonian equation $\nabla^2\varphi = 4\pi\,G\rho$. Note that it is $(\rho+3p)$, rather than just ρ, which appears on the right in (86). We may regard $(\rho+3p)$ as representing the "active gravitational mass density" (i.e., the quantity which describes the ability to make gravitational field) of the fluid.

Further support for this interpretation comes from a simple calculation. Set $\eta^a = (-\lambda)^{-1/2}t^a$ (so η^a is the unit four-velocity of an observer following an integral curve of t^a, the time-translation). Then

$$\eta^b\nabla_b\eta^a = \eta^b\nabla_b[(-\lambda)^{-1/2}t^a] = t^a\eta^b\nabla_b(-\lambda)^{-1/2} + (-\lambda)^{-1/2}\eta^b\nabla_b t^a$$
$$= 0 + (-\lambda)^{-1/2}\eta^b\nabla_b t^a = -\lambda^{-1}t^b\nabla_b t^a = \lambda^{-1}t^b\nabla^a t_b \tag{87}$$
$$= \frac{1}{2}\lambda^{-1}\nabla^a(t^bt_b) = \frac{1}{2}\lambda^{-1}\nabla^a\lambda = \nabla^a[\frac{1}{2}\ln(-\lambda)]$$

Thus, the acceleration felt by this observer is $\nabla^a[\frac{1}{2}\ln(-\lambda)]$. Eqn. (87) is analogous to the Newtonian equation $\vec{A} = -\vec{\nabla}\varphi$ for the acceleration felt by a free particle in potential φ. Why the difference in sign between the Newtonian expression and (87)? Because of the difference in the meaning of "acceleration" in the two cases. In the Newtonian case, we release the particle, and observe its acceleration (relative to, e.g., the star) toward the star. In general relativity, such a particle would be said to have zero acceleration (since acceleration is measured by the cube and springs). Instead we hold the particle at fixed distance from the star, and then note its acceleration (measured by the cube). In the Newtonian case, the free particle "accelerates" toward the star; in general relativity, the held particle is (since it is held) accelerated away from the star.

The remarks above are intended to motivate and to give a feeling for the calculations. We now discuss the geometry in detail.

We first obtain some additional facts about the Killing vectors. We have

$$\underset{1}{l}^a\nabla_a\lambda = \underset{1}{l}^a\nabla_a(t^bt_b) = 2t^b\underset{1}{l}^a\nabla_a t^b = 2t^bt^a\nabla_a\underset{1}{l}_b = 0$$

where, in the second step, we have expanded using the Leibniz rule, in the third step, we have used $[\underset{1}{l}^a, t]^a = 0$, and, in the fourth step, we have used $\nabla_{[a}\underset{1}{l}_{b]} = 0$. In words, λ *is invariant under the rotations.* (This, of course, is what we would expect.) We next prove that *the rotational Killing vectors are orthogonal to the timelike Killing vector.* The proof is straightforward:

$$\begin{aligned} t^a\underset{3}{l}_a &= t^a(\underset{1}{l}^b\nabla_b\underset{2}{l}_a - \underset{2}{l}^b\nabla_b\underset{1}{l}_a) = -t^a\underset{1}{l}^b\nabla_a\underset{2}{l}_b + t^a\underset{2}{l}^b\nabla_a\underset{1}{l}_b \\ &= -\underset{2}{l}^a\underset{1}{l}^b\nabla_a t_b + \underset{1}{l}^a\underset{2}{l}^b\nabla_a t_b = 2\underset{1}{l}^a\underset{2}{l}^b\nabla_a t_b \\ &= 2\underset{1}{l}^a\underset{2}{l}^b(-\lambda^{-1}t_{[a}\nabla_{b]}\lambda) = 0 \end{aligned}$$

where, in the first step we have used $\underset{3}{l}_a = [\underset{1}{l}, \underset{2}{l}]_a$; in the second, Killing's equation on $\underset{1}{l}_a$ and $\underset{2}{l}_a$; in the third, $[\underset{1}{l}, t]^a = [\underset{2}{l}, t]^a = 0$; in the fourth, Killing's equation on t_a; in the fifth, (81); and, in the sixth step, the fact that $\underset{1}{l}^a\nabla_a\lambda = \underset{2}{l}^a\nabla_a\lambda = 0$. It is perhaps not surprising that "rotation" gives a motion in a space-time orthogonal to the motion "time-translation".

We introduce a (positive) scalar field r in space-time:

$$2r^2 = \underset{1}{l}^a\underset{1}{l}_a + \underset{2}{l}^a\underset{2}{l}_a + \underset{3}{l}^a\underset{3}{l}_a \tag{88}$$

We interpret r as representing a sort of "radial distance from the center of the star", an interpretation justified by the observation that this is precisely what the r given by (83) is for rotational Killing vectors in Euclidean space. We next show that r *is invariant under the time-translation and the rotations* (more precisely, $t^a\nabla_a r = \underset{1}{l}^a\nabla_a r = \underset{2}{l}^a\nabla_a r = \underset{3}{l}^a\nabla_a r = 0$) (Note, again, that this assertion is obvious geometrically.) First, the time-translation part:

$$t^a\nabla_a(\underset{1}{l}^b\underset{1}{l}_b) = 2t^a\underset{1}{l}^b\nabla_a\underset{1}{l}_b = 2\underset{1}{l}^a\underset{1}{l}^b\nabla_a t_b = 0$$

where, in the second step, we have used $[t, \underset{1}{l}]^a = 0$. repeating for $\underset{2}{l}^b\underset{2}{l}_b$ and $\underset{3}{l}^b\underset{3}{l}_b$, and summing, using (88), we obtain $t^a\nabla_a r = 0$. For the rotation part:

$$\begin{aligned} \underset{1}{l}^a\nabla_a(2r^2) &= \underset{1}{l}^a\nabla_a(\underset{1}{l}^b\underset{1}{l}_b + \underset{2}{l}^b\underset{2}{l}_b + \underset{3}{l}^b\underset{3}{l}_b) \\ &= 2\underset{1}{l}^a\underset{1}{l}^b\nabla_a\underset{1}{l}_b + 2\underset{1}{l}^a\underset{2}{l}^b\nabla_a\underset{3}{l}_b + 2\underset{1}{l}^a\underset{3}{l}^b\nabla_a\underset{2}{l}_b \\ &= 0 + 2\underset{2}{l}^b(\underset{2}{l}^a\nabla_a\underset{1}{l}_b + \underset{3}{l}_b) + 2\underset{3}{l}^b(\underset{3}{l}^a\nabla_a\underset{1}{l}_b - \underset{2}{l}_b) \\ &= 2\underset{2}{l}^b\underset{3}{l}_b - 2\underset{3}{l}^b\underset{2}{l}_b = 0 \end{aligned}$$

where, in the third step, we have used (79), and, in the fourth step, we have used Killing's equation on $\underset{1}{l}_b$.

Finally, we introduce the scalar field $\mu = (\nabla^a r)\nabla_a r$. In flat space, $\mu = 1$ (since r is distance). Hence, deviations of μ from this value represent "curvature of space". By an argument similar to those above (or, using properties of Lie

derivatives, in one step) we have that μ *is invariant under the time-translation and rotations.*

Let us summarize the situation. We think of r as a "radial coordinate". We think of λ and μ as "fields which describe the geometry of space-time". Since our space-time is static and spherically symmetric, we expect that everything of interest will be a function only of r. This expectation is made precise as follows. We have seen that the derivative of any of r, μ or λ in the direction of any of t^a, $\underset{1}{l}^a$, $\underset{2}{l}^a$, or $\underset{3}{l}^a$ is zero. Since t^a, $\underset{1}{l}^a$, $\underset{2}{l}^a$, and $\underset{3}{l}^a$ are linearly dependent, and t^a is orthogonal to all three, these four vectors span a three-dimensional vector space at each point of space-time. Hence, there is only one direction, at each point, simultaneously orthogonal to t^a, $\underset{1}{l}^a$, $\underset{2}{l}^a$, and $\underset{3}{l}^a$. But $\nabla_a r$, $\nabla_a \lambda$ and $\nabla_a \mu$ are simultaneously orthogonal to these vectors. Hence, $\nabla_a r$, $\nabla_a \lambda$ and $\nabla_a \mu$ are all proportional to each other. In other words, $\lambda = \lambda(r)$ and $\mu = \mu(r)$, functions of r.

The idea is to use Einstein's equation to obtain a pair of ordinary differential equations on the functions $\lambda(r)$ and $\mu(r)$ We have one such equation, (85). Let us work in the region outside the star ($\rho = p = 0$). Then (85) becomes

$$\nabla^2 \lambda - \lambda^{-1} (\nabla^a \lambda)(\nabla_a \lambda) = 0 \tag{89}$$

We need a second equation. From the definition of the Riemann tensor, $\nabla_{[a} \nabla_{b]} \nabla_c r = \frac{1}{2} R_{abc}{}^m \nabla_m r$. Contracting over "a" and "c" and then contracting with $\nabla^b r$ we obtain

$$\begin{aligned} &\frac{1}{2} \nabla^2 (\nabla^a r \nabla_a r) - (\nabla^a \nabla^b r)(\nabla_a \nabla_b r) - \nabla^a r \nabla_a (\nabla^2 r) \\ &= R_{ab} \nabla^a r \nabla^b r = 4\pi \, G \, \mu (\rho - p) \end{aligned}$$

where we have eliminated the Ricci tensor using (84). Hence,

$$\frac{1}{2} \nabla^2 (\nabla^a r \nabla_a r) - (\nabla^a \nabla^b r)(\nabla_a \nabla_b r) - \nabla^a r \nabla_a \nabla^2 r \tag{90}$$

Here is our second piece of Einstein's equation. All that remains is to rewrite (89) and (90) as ordinary differential equations on $\lambda(r)$ and $\mu(r)$.

Denote d/dr by a prime. Then $\nabla_a \lambda = \lambda' \nabla_a r$ (chain rule). Hence, $\nabla^2 \lambda = (\nabla^a \lambda') \nabla_a r + \lambda' \nabla^2 r = \lambda'' \nabla^a r \nabla_a r + \lambda' \nabla^2 r = \lambda'' \mu + \lambda' \nabla^2 r$. Similarly, $\nabla^2 \mu = \mu'' \mu + \mu' \nabla^2 r$. Hence, (89) and (90) become

$$\lambda'' \mu + \lambda' \nabla^2 r - \lambda^{-1} (\lambda')^2 \mu = 0 \tag{91}$$

$$\frac{1}{2} \mu \mu'' + \frac{1}{2} \mu' \nabla^2 r - (\nabla^a \nabla^b r)(\nabla_a \nabla_b r) - \mu (\nabla^2 r)' = 0 \tag{92}$$

respectively.

Eqns. (91) and (92) are still not ordinary differential equations: they contain derivatives of r. The final step is to evaluate $\nabla_a \nabla_b r$ for substitution into (91) and (92). We have:

$$\begin{aligned} t^b \nabla_a \nabla_b r &= \nabla_a (t^b \nabla_b r) - (\nabla^b r) \nabla_a t_b = 0 - (\nabla^b r)[-\lambda^{-1} t_{[a} \nabla_{b]} \lambda] \\ &= \frac{1}{2} \lambda^{-1} t_a (\nabla^b r)(\nabla_b \lambda) = \frac{1}{2} \lambda^{-1} \lambda' \mu \, t_a \end{aligned}$$

and

$$(\nabla^b r) \nabla_a \nabla_b r = \frac{1}{2} \nabla_a (\nabla^b r \nabla_b r) = \frac{1}{2} \mu' \nabla_a r$$

Finally,

$$\begin{aligned}\underset{1}{l}^a\underset{1}{l}^b\nabla_a\nabla_b r &= \underset{1}{l}^a\nabla_a(\underset{1}{l}^b\nabla_b r) - \underset{1}{l}^a(\nabla^b r)\nabla_a\underset{1}{l}_b \\ &= 0 + \underset{1}{l}^a(\nabla^b r)\nabla_b\underset{1}{l}_a = \frac{1}{2}(\nabla^b r)[\nabla_b(\underset{1}{l}^a\underset{1}{l}_a)]\end{aligned}$$

and similarly for $\underset{3}{l}^a\underset{2}{l}^b$ and $\underset{3}{l}^a\underset{3}{l}^b$. Hence,

$$\begin{aligned}&(\underset{1}{l}^a\underset{1}{l}^b + \underset{2}{l}^a\underset{2}{l}^b + \underset{3}{l}^a\underset{3}{l}^b)\nabla_a\nabla_b r = \frac{1}{2}(\nabla^b r)\nabla_b(\underset{1}{l}^a\underset{1}{l}_a + \underset{2}{l}^a\underset{2}{l}_a + \underset{3}{l}^a\underset{3}{l}_a) \\ &= \frac{1}{2}(\nabla^b r)\nabla^b(2r^2) = 2r(\nabla^b r)(\nabla_b r) = 2r\mu\end{aligned}$$

We now have expressions for the components of $\nabla_a\nabla_b r$ in the t^a-direction, the $\nabla^a r$-direction, and the $\underset{i}{l}^a\underset{i}{l}^b$ directions. Hence,

$$\begin{aligned}\nabla_a\nabla_b r = &\frac{1}{2}\lambda^{-2}\lambda'\mu t_a t_b + \frac{1}{2}\mu^{-1}\mu'(\nabla_a r)(\nabla_b r) \\ &+ \mu r^{-1}[g_{ab} - \lambda^{-1}t_a t_b - \mu^{-1}(\nabla_a r)(\nabla_b r)]\end{aligned} \tag{93}$$

Therefore,

$$\nabla^2 r = \frac{1}{2}\lambda^{-1}\lambda'\mu + \frac{1}{2}\mu' + 2\mu r^{-1}$$

$$(\nabla^a\nabla^b r)(\nabla_a\nabla_b r) = \frac{1}{4}\lambda^{-2}\mu^2(\lambda')^2 + \frac{1}{4}(\mu')^2 + 2\mu^2 r^{-2}$$

Substituting, (91) and (92) become, respectively,

$$\lambda''\mu - \frac{1}{2}\lambda^{-1}\mu\,(\lambda')^2 + \frac{1}{2}\lambda'\mu' + 2\mu\, r^{-1}\lambda' = 0 \tag{94}$$

$$-\frac{1}{4}\lambda^{-1}\mu\,\lambda'\mu' - \mu\,\mu' r^{-1} + \frac{1}{4}\lambda^{-2}\mu^2(\lambda')^2 - \frac{1}{2}\lambda^{-1}\mu^2\lambda'' = 0 \tag{95}$$

We have now obtained the ordinary differential equations we sought. What remains is to solve them. Eliminating λ'' between (94) and (95), we obtain simply $\lambda'/\lambda = \mu'/\mu$. So, λ is a constant multiple of μ. What multiple should we choose?

In Minkowski space, $\lambda = -1$, and $\mu = 1$, which suggests $\lambda = -\mu$. In particular, if we want things to be scaled properly asymptotically (far from the star, where space-time is nearly flat) we should make this choice. We do so. Setting $\lambda = -\mu$ in (95), we obtain immediately $\lambda''/\lambda' + 2r^{-1} = 0$. The solution is $\lambda = a + b/r$, where a and b are constants. What should we pick? Asymptotically, (i.e., for large r), we want λ to approach -1 (that is, we wish to so scale t^a). So, we choose $a = -1$. We write $\lambda = -1 + 2GM/r$ where G is the gravitational constant, and M is another constant. We wish to interpret M. Let us suppose we are far from the star, so $2Gm/r \ll 1$ (we shall discuss this issue in a moment). Then, to first order in $2GM/r$, $\frac{1}{2}\ln(-\lambda) = \frac{1}{2}\ln(1 - 2GM/r) = -GM/r$. But $\frac{1}{2}\ln(-\lambda)$ is, as we have seen, the quantity analogous to the Newtonian potential, while $-GM/r$ is the potential, in Newtonian theory, of a point mass of mass M. Hence, we interpret the number M as the "apparent mass of the star, as seen by its distinct gravitational effects". This remark is made more explicit by (85). Clearly, from this equation there follows an expression (which is not very interesting) for M as an integral (of $(\rho + p)$) over the star.

It should now be clear that one can choose coordinates in which the metric for the Schwarzschild solution takes the well-known form

$$-(1-\frac{2GM}{r})\,\mathrm{d}t^2+(1-\frac{2GM}{r})^{-1}\,\mathrm{d}r^2+r^2(\,\mathrm{d}\theta^2+\sin^2\theta\,\mathrm{d}\varphi^2)$$

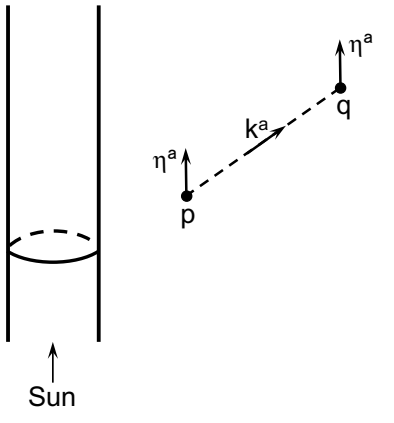
The θ and φ are "angular coordinates", while the scalar field r becomes a "radial coordinate".

Finally, we remark on the sign-reversal of $(1-2GM/r)$. when $r=r_c=2GM$, this expression (for λ and μ) reverses sign. This r_c associated with any mass is called the *Schwarzschild radius* of that mass. (For the Sun, r_c is a mile or so; for the Earth, a millimeter.) Most bodies, of course, are larger then their Schwarzschild radius, so λ is always negative and μ positive. In collapse phenomenon, however, a body passes inside its Schwarzschild radius, to a region of space-time in which t^a becomes spacelike. Clearly, the approximation $2GM/r \ll 1$ is reasonable in most situations. Physically, r_c for a body is approximately the size that body must be in order that the Newtonian escape velocity equal the speed of light. Thus, "light cannot escape" from a body inside its Schwarzschild radius. This is reflected, geometrically, by a "turning in of the light cones", and, in the formalism, by the sign reversals of λ and μ.

26. The Schwarzschild Solution: Observations

Presumably, the Schwarzschild solution represents a good approximation to the geometry of space-time in the vicinity of our solar system. The fluid mass at the center is the Sun; the planets are "free test particles" which follow geodesics. Does this general relativistic model of our solar system agree in its broad features with the (very accurate) model from Newtonian gravitation? Are there any new, observable consequences of the model in general relativity? The answer to both questions is yes. The new features of the relativistic model are the three famous "tests" of general relativity.

We begin with the red-shift. The light we see from the surface of the sun is observed with a longer wavelength than the light had on leaving the surface. What is the prediction, from the Schwarzschild solution, of this effect? Of course the frequency (and wavelength) of a beam of light depends on a four-velocity of he who observes it. To discuss red-shift, therefore, we must decide who is looking at the light. Let us consider observers who are "static", i.e., whose four-velocities are proportional to the timelike Killing vector t^a.



The four-velocity of such an observer is $\eta^a = (-\lambda)^{-1/2} t^a$ (since $t^a t_a = \lambda < 0$). Now consider a light ray with tangent vector k^a. Then the frequency of the light, as seen by our observer, is $\omega = -\eta^a k_a = -(-\lambda)^{-1/2} t^a k_a$. Let this light ray pass from point p to point q of space-time. Then, since $k_a t^a$ is constant along the geodesic, we have

$$\frac{\omega(q)}{\omega(p)} = \left(\frac{\lambda(p)}{\lambda(q)}\right)^{1/2} = \left(\frac{1 - 2GM/r(p)}{1 - 2GM/r(q)}\right)^{1/2}$$

where we have used $\lambda = -(1 - 2GM/r)$. This is the formula for the red-shift. It enables us to compute the change in the frequency of light (as seen by an observer moving with the timelike Killing vector) as it passes from any point of space-time to any other point. In practice, one uses this equation in the case $2GM/r \ll 1$ (always such in the solar system, where $2GM$ is a kilometer). Then, to first order in $2GM/r$, the equation above becomes

$$\frac{\omega(q)}{\omega(p)} = 1 - \frac{GM}{r(p)} + \frac{GM}{r(q)} \qquad (GM \ll r)$$

In particular, for $r(q) > r(p)$, $\omega(q) < \omega(p)$, so the apparent wavelength of the light indeed increases as the light goes outward from the Sun. Solar red-shift experiments are good to about 10%, and agree with the theory. A much more accurate (tenths of a percent) red-shift experiment has been performed using the Earth, again, with, agreement with the theory.

The discussion of the other two tests requires more details about the

behavior of geodesics. We now derive what we shall need. Consider a geodesic with tangent vector p^a, and fix a point q on this geodesic. Draw the two-sphere (of spherical symmetry) through q. Then we can choose the rotational Killing vectors so that $\underset{3}{l}^a$ vanishes at q, $\underset{1}{l}^a$ is parallel to the projection of p^a onto the sphere, and $\underset{2}{l}^a$ is orthogonal to the projection. (That such choices are possible is clear from the geometry of a sphere. Each Killing vector vanishes at a pair of antipodal points (the "axis of the rotation"). Choose $\underset{3}{l}^a$ to vanish at q, and $\underset{1}{l}^a$ and $\underset{2}{l}^a$ to vanish on the equator, where q is the North pole. Finally, let the projection of p^a into the sphere point along the meridian which goes to the point where $\underset{2}{l}^a$ vanishes.) At the point q, $p_a\underset{2}{l}^a = p_a\underset{3}{l}^a = 0$. Hence, this remains true along the entire geodesic. Therefore, it is true along the entire geodesic that $\underset{1}{l}^a\underset{1}{l}_a = r^2$. (At q, for example $\underset{1}{l}^a\underset{1}{l}_a = \underset{2}{l}^a\underset{2}{l}_a$, and $\underset{3}{l}^a\underset{3}{l}_a = 0$. Hence, from (88) $\underset{1}{l}^a\underset{1}{l}_a = r^2$.)

From the choice above for Killing vectors, we have

$$p^a = \kappa t^a + \sigma \underset{1}{l}^a + \tau \nabla^a r \tag{96}$$

where κ, σ, and τ are functions along the geodesic. To evaluate them, we use the fact that $E = -p_a t^a = -(\kappa t_a + \sigma \underset{1}{l}_a + \tau \nabla_a r)t^a = -\kappa\lambda$ is constant along the geodesic, and $L = p_a \underset{1}{l}^a = r^2\sigma$ is constant along the geodesic. Hence,

$$p^a = -E\,\lambda^{-1}t^a + L\,r^{-2}\underset{1}{l}^a + \tau\nabla^a r \tag{97}$$

where E and L are constants (the energy and angular momentum, respectively). To determine τ, we use the fact that $p^a p_a = -m^2$ is constant along the geodesic. Hence,

$$E^2\lambda^{-1} + L^2 r^{-2} + \tau^2\mu = -m^2 \tag{98}$$

In (98), E, L and m are constants, while $-\lambda = \mu = (1 - 2GM/r)$. Hence, everything is known but τ, so (98) gives τ as a function of r. Thus, from symmetry considerations alone, we determine the tangent vector to our geodesic (97), given the values of the energy E, the angular momentum L, and the rest mass m.

The next step is to use (97) and (98) to determine the rate of change of certain quantities of interest along the geodesic. The first such quantity is r. We have $p^a\nabla_a r = (-E\,\lambda^{-1}t^a + L\,r^2\underset{1}{l}^a + \tau\nabla^a r)\nabla_a r$. Since t^a and $\underset{1}{l}^a$ are both orthogonal to $\nabla_a r$ we have

$$p^a\nabla_a r = \tau\mu = \mu^{1/2}[-E^2\,\lambda^{-1} - m^2 - L^2\,r^{-2}]^{1/2} \tag{99}$$

were, in the second step, we have used (98). The other "quantity of interest" is an angular function φ, defined by $\nabla^a r\nabla_a\varphi = t^a\nabla_a\varphi = 0$, $\underset{1}{l}^a\nabla_a\varphi = 1$ (i.e., defined by writing down the properties of the angular coordinate φ in Euclidean space). From (97),

$$p^a\nabla_a\varphi = L\,r^{-2} \tag{100}$$

The equations we wanted are (99) and (100). Note, from (100), that φ is strictly increasing along the geodesic. Hence, we may take φ as a parameter along our

geodesic, and take r along the geodesic as a function of this parameter φ. To find the differential equation for $r(\varphi)$, divide (99) by (100):

$$\frac{\mathrm{d}r}{\mathrm{d}\varphi} = r^2 L^{-1}\,\mu^{1/2}[-E^2\,\lambda^{-1} - m^2 - L^2\,r^{-2}]^{1/2} \tag{101}$$
$$= r^2 L^{-1}\left(\frac{1-2GM}{r}\right)^{1/2}\left[E^2\left(1-\frac{2GM}{r}\right)^{-1} - m^2 - L^2\,r^{-2}\right]^{1/2}$$

The right side of (101) is an explicit function of r. Hence, (101) can, in principle, be integrated to give $r(\varphi)$. The problem of finding the geodesics in the Schwarzschild solution reduces to that of performing a certain integration. Unfortunately, the result is an elliptic integral, whose properties are not particularly transparent. We therefore proceed to analyze (101) in a more indirect way. First, replace r in (101) by $u = 1/r$, to obtain

$$\frac{\mathrm{d}u}{\mathrm{d}\varphi} = -L^{-1}[E^2 - m^2(1-2GM\,u) - L^2\,u^2(1-2GM\,u)]^{1/2}$$

Next, square this equation, and take $\mathrm{d}/\,\mathrm{d}\varphi$ of both sides:

$$2\frac{\mathrm{d}u}{\mathrm{d}\varphi}\frac{\mathrm{d}^2u}{\mathrm{d}\varphi^2} = \frac{2GM\,m^2}{L^2}\frac{\mathrm{d}u}{\mathrm{d}\varphi} - 2u\frac{\mathrm{d}u}{\mathrm{d}\varphi} + 6GM\,u^2\frac{\mathrm{d}u}{\mathrm{d}\varphi}$$

Finally, rearrange terms:

$$\frac{\mathrm{d}^2u}{\mathrm{d}\varphi^2} + u = \frac{GM\,m^2}{L^2} + 3GM\,u^2 \tag{102}$$

Clearly, every solution of (101) is also a solution of (102). It should also be clear that (102) is easier to think about than (101). Our approach to (102) will be based on an approximation procedure involving $GM/r \ll 1$, which, as we have seen, is well satisfied in our solar system.

We first consider the case $m > 0$, i.e., the case of orbits of material particles. To zeroth order in GM/r, we may ignore the term $3GM\,u^2$ on the right in (102), for it is negligible compared with u. Then (102) becomes $\dfrac{\mathrm{d}^2u}{\mathrm{d}\varphi^2} + u = GM\,m^2/L^2$, with solution $u = GM\,m^2\,L^{-2} + GM\,m^2\,L^{-2}\epsilon\,\cos\varphi$, where ϵ is a constant. Expressed in terms, of r,

$$r = \frac{L^2}{GM\,m^2}\frac{1}{1+\epsilon\,\cos\varphi} = r_0\frac{1}{1+\epsilon\,\cos\varphi} \tag{103}$$

Eqn. (103) will be recognized as the equation for an ellipse. Thus, we obtain, to zeroth order in GM/r, the usual elliptical Newtonian orbits. Note in particular that, to this order, $u(\varphi)$ is periodic in φ with period 2π.

The relativistic perturbations from the Newtonian results are obtained by considering the next order term in GM/r. We seek a solution of (102) of the form

$$u = GM\,m^2\,L^{-2}(1+\epsilon\,\cos\varphi) + \delta u \tag{104}$$

where $\delta u \ll u$. Substituting, and keeping first order terms, we obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}\varphi^2}\delta u + \delta u = 3\,G^3M^3m^4L^{-4}(1+\epsilon\,\cos\varphi)^2 \tag{105}$$

This is the equation for the perturbation, δu from the Newtonian orbit. But (105) is easy to solve:

$$\delta u = 3\,G^3M^3m^4L^{-4}[1+\frac{1}{2}\epsilon^2+\frac{1}{6}\epsilon^2\cos 2\varphi+\epsilon\,\varphi\,\sin\varphi] \tag{106}$$

Hence, to first order in GM/r, the solution of (102) is

$$\begin{aligned} u &= GM\,m^2\,L^{-2}(1+\epsilon\,\cos\varphi) \\ &\quad + 3\,G^3M^3m^4L^{-4}[1+\frac{1}{2}\epsilon^2+\frac{1}{6}\epsilon^2\cos 2\varphi+\epsilon\,\varphi\,\sin\varphi] \end{aligned} \tag{107}$$

What is the effect of the perturbation term in (107)? The terms $1+\frac{1}{2}\epsilon^2+\frac{1}{6}\epsilon^2\cos 2\varphi$ are periodic in φ with period 2π. These terms change the "shape" of the orbit, but do not cause the ellipse to precess. The precession is caused by the term $\epsilon\,\varphi\,\sin\varphi$ on right in (107). To see its effect, let us ignore" the shape-changing terms", $1+\frac{1}{2}\epsilon^2+\frac{1}{6}\epsilon^2\cos 2\varphi$, on the right in (107) (these are unobservable.) Then (107) becomes

$$u = GM\,m^2\,L^{-2}+\epsilon\,GM\,m^2\,L^{-2}\,\cos\varphi+3\,\epsilon\,G^3M^3m^4L^{-4}\varphi\;sin\varphi \tag{108}$$

But, for $\kappa \ll 1$, $\cos(1+\kappa)\varphi = \cos(\kappa\varphi)\cos\varphi - \sin(\kappa\varphi)\sin\varphi = \cos\varphi - \kappa\varphi\sin\varphi$. Hence, to first order in GM/r, (108) gives

$$u = GM\,m^2\,L^{-2}[1+\epsilon\,\cos((1-3\,G^2M^2\,m^2\,L^{-2})\varphi)] \tag{109}$$

The situation should now be clear. The elliptical Newtonian orbit of the planet precessed with each revolution of the planet. For each orbit, the axis of the ellipse moves through angle

$$\Delta\varphi = 6\,\pi\,G^2M^2\,m^2\,L^{-2} = 6\,\pi\,G\,M/r_0 \tag{110}$$

where, in the second expression, we have set $r_0 = L^2/GM\,m^2$ (see (103)). Eqn, (110) is the equation for the precession of the perihelion of a planetary orbit, so the precession effect is first order in GM/r_0.

To summarize, general relativity causes "corrections" in the Newtonian planetary orbits. The orbit changes its shape (which is unobservable), and, in addition, the perihelion (point of closest approach of the planet to the Sun) precesses, for each orbit, through an angle given by (110). For the planet Mercury, such a precession has been observed (the angle is about 10^{-7} radians per orbit), and agrees with the relativistic prediction to within about 5%.

Finally, we consider the geodesic equation, (102), in the case $m = 0$. Thus, we consider the behavior of light rays. This yields the third classical test of general relativity. With $m = 0$, (102) becomes

$$\frac{\mathrm{d}^2u}{\mathrm{d}\varphi^2}+u = 3\,G\,M\,u^2 \tag{111}$$

Again, we solve (111) by an approximation procedure, using $GM/r \ll 1$. Ignoring the small right side of (111) (i.e., working to zeroth order in GM/r) we have the solution

$$u = \frac{1}{r_0}\cos\varphi \tag{112}$$

This, of course, is the equation for a straight line in polar coordinates. The first order term in GM/rchanges this curve slightly. To find the change, we set

$$u = \frac{1}{r_0}\cos\varphi+\delta u \tag{113}$$

with $\delta u \ll u$. Substituting into (111), and keeping first order terms,

$$\frac{\mathrm{d}^2 \delta u}{\mathrm{d}\varphi^2} + \delta u = \frac{3\,G\,M}{{r_0}^2} \cos^2 \varphi \tag{114}$$

The solution of (114) is $\delta u = (GM/{r_0}^2)(2 - \cos^2 \varphi)$. Hence, the solution of (111), good to first order in GM/r, is

$$u = \frac{1}{r_0} \cos \varphi + \frac{2\,G\,M}{{r_0}^2} - \frac{G\,M}{{r_0}^2} \cos^2 \varphi \tag{115}$$

Here, we are concerned with the asymptotic behavior of the light ray, i.e., its behavior far from the star. But r approaches infinity when u $(= 1/r)$ approaches zero. Hence, we are interested in the φ-values for which the right side of (115) vanishes. To first order in GM/r, the right side of (115) vanishes for $\cos \varphi = -2\,GM/r_0$, or for $\varphi = \frac{\pi}{2} + \frac{2GM}{r_0}$ and $-\frac{\pi}{2} - \frac{2GM}{r_0}$. Clearly, the light "bends" as it passes the Sun, through an angle

$$\Delta\varphi = \frac{4\,G\,M}{r_0} \tag{116}$$

This is the equation for the deflection of light passing by a body in general relativity. Note that r_0 on the right in (116) is the distance of the light from the Sun at the point of closest approach.

To summarize, light passing by a body is bent by the presence of that body, through an angle given by (116). This effect has been observed, by looking at light from distant stars passing close to the Sun. The observations agree with general relativity to perhaps 25%.

In interpreting the last two tests, we have implicitly assumed that, when an effect was expressed in terms of r and φ, then this is the effect that should be observed, using Euclidean coordinates in our solar system. This assumption should, strictly speaking, be justified by a more detailed analysis of what is actually observed. Such a justification is possible, but rather tedious. Suffice it to say that, for effects to first order in GM/r, r and φ adequately describe the "Euclidean geometry" with respect to which observationalists would orient themselves to report the results of their experiments.

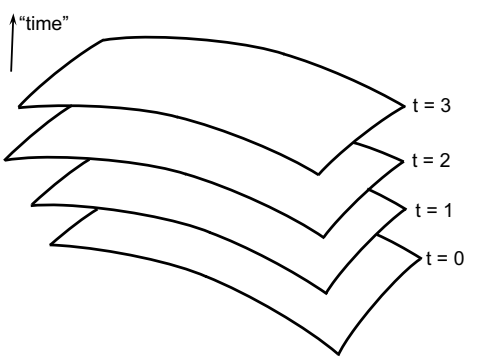
27. The Initial-Value Formulation

An essential concept in physics is that of prediction. Given sufficient detail about what is happening now, one predicts, using the equation which describe the system, what will happen later. This power of prediction is one of the important things which makes a physical theory a physical theory. It is available in essentially every physical theory. That prediction (of the future from the present) is indeed possible within the context of a given theory is made manifest by casting that theory into initial-value form. The initial-value formulation of a theory normally has the following appearance. There is a certain set of information about the system (at any one instant of time) called the data for the system. The theory gives a set of differential equations of the form $\mathrm{d}/\mathrm{d}t(\text{data}) = F(\text{data})$, where $\mathrm{d}/\mathrm{d}t$ is time derivative, and F is some function of the data. These are the evolution equations. If we know the data now, the evolution equations determine the data an instant later, then the data another instant later, etc. In this way, one predicts.

Of course, the notions above are not natural ones in general relativity. They refer to "time", and "things happening". In general relativity, there is only space-time: there is no time, no evolution – nothing happens. It is of interest, nonetheless, to ask whether general relativity can be forced into initial-value formulation. The answer is that, if one is willing to sacrifice some of the beauty, it can.

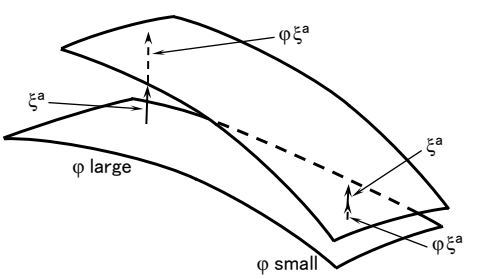
We wish to speak of time. Hence, we introduce a scalar field t on space-time with $\nabla_a t$ timelike. The various surfaces $t = \text{const}$ (one surface for each value of "const") thus represent surface of constant time. These are three-dimensional spacelike surfaces, and they fill space-time. Of course, a given space-time can be "sliced" as below in many different ways. This reflects the fact that "time" is not natural in general relativity. In order to introduce time, we must introduce additional structure (the t above) into space-time.

The initial-value formulation of general relativity results from recasting Einstein's equation in the presence of a scalar field t with timelike gradient. We first derive the equations, and then, at the end, discuss their significance.



Set $\varphi = [-(\nabla^a t)(\nabla_a t)]^{-1/2}$, and $\xi^a = -\varphi \nabla^a t$. Thus, ξ^a is a unit ($\xi^a \xi_a = -1$) vector field in the direction of $\nabla^a t$. That is, ξ^a is the unit normal vector field to our family of surfaces. Furthermore, $\varphi \xi^a \nabla_a t = 1$. Thus, we may interpret $\varphi \xi^a$ as the "connecting vector" between nearby surfaces. If, for example, φ were constant, then all the surfaces would be equidistant from their (infinitesimally nearby) neighbors. For large φ, the surfaces are far apart; for small φ, close together.

A tensor field in space-time will be called *spatial* (with respect to the "field of observers" defined by ξ^a) if any index of that tensor field, contracted with ξ^a, gives zero. In particular, $h_{ab} = g_{ab} + \xi_a \xi_b$ is a spatial tensor field, the spatial metric. As usual, $h_a{}^m h_{bm} = h_{ab}$, so h_{ab} is also a spa-



tial projection tensor. Finally, we introduce a dot for $\pounds_{\varphi\xi}$. Recalling that $\varphi\xi^a$ is the connecting vector between nearby surfaces, we can think of $\pounds_{\varphi\xi}$ as the "time derivative". We shall also need the spatial derivative operator, defined by

$$D_a T_{b\ldots c} = h_a{}^m h_b{}^n \cdots h_c{}^p \nabla_m T_{n\ldots p} \tag{117}$$

on any spatial tensor field. That this is the correct spatial derivative can be verified by checking that $D_a h_{bc} = 0$ (the defining property of the derivative).

The idea is to introduce certain tensor fields on each surface t = constant to represent the data. Then, we wish to obtain equations giving the "dot" of each data tensor field in terms of the values of those fields. In this way, we acquire the capability of integrating – evolving – from one surface to the next.

The first tensor field is the spatial metric h_{ab}. The second is a field Π called the extrinsic curvature, and defined by

$$\Pi^{ab} = \nabla^a \xi^b + \varphi^{-1} \xi^a \nabla^b \varphi + \varphi^{-1} \xi^a \xi^b (\xi^m \nabla_m \varphi) \tag{118}$$

We first show that this Π^{ab} is a symmetric, spatial tensor field. To show symmetry, antisymmetrize (118) over "a" and "b":

$$\Pi^{[ab]} = \nabla^{[a} \xi^{b]} + \varphi^{-1} \xi^{[a} \nabla^{b]} \varphi = \varphi \nabla^{[a} (\varphi^{-1} \xi^{b]}) = 0 \tag{119}$$

where, in second step, we have combined terms, and, in the third, we have used the fact that $\varphi^{-1}\xi_a = -\nabla_a t$ whence the curl of this vector field vanishes. To see that Π^{ab} is spatial, contract with ξ_b

$$\begin{aligned} \xi_b \Pi^{ab} &= \xi_b \nabla^a \xi^b + \varphi^{-1} \xi^a \xi_b \nabla^b \varphi + \varphi^{-1} \xi_b \xi^a \xi^b (\xi^m \nabla_m \varphi) \\ &= 0 + \varphi^{-1} \xi^a (\xi^b \nabla_b \varphi) - \varphi^{-1} \xi^a (\xi^m \nabla_m \varphi) = 0 \end{aligned} \tag{120}$$

Since $\Pi^{ab} = \Pi^{ba}$, it follows also that $\xi_a \Pi^{ab} = 0$.

The data for general relativity are the two tensor fields and h_{ab} and Π^{ab}. We think of h_{ab} as analogous to "x" (position of a particle) in particle mechanics, and Π^{ab} as analogous to p (momentum of the particle). If you are given the position and momentum of a particle initially, the future motion of the particle is determined. Similarly, if you are given the "configuration of the geometry" (the spatial metric h_{ab}) and its "momentum" (the extrinsic curvature Π^{ab}), the space-time is determined. We now derive the equations which describe how this "determination" comes about.

We first consider h_{ab}:

$$\begin{aligned} &h'_{ab} = \pounds_{\varphi\,\xi} h_{ab} = h_a{}^m h_b{}^n \pounds_{\varphi\,\xi} h_{mn} \\ &= h_a{}^m h_b{}^n [\varphi\, \xi^p \nabla_p h_{mn} + h_{mp} \nabla_n (\varphi\, \xi^p) + h_{pn} \nabla_m (\varphi\, \xi^p)] \\ &= h_a{}^m h_b{}^n [\varphi\, \xi^p \nabla_p (g_{mn} + \xi_m \xi_n) + \varphi\, h_{mp} \nabla_n \xi^p + \varphi\, h_{pn} \nabla_m \xi^p] \\ &= 2\, \varphi\, \Pi_{ab} \end{aligned} \tag{121}$$

where, in the second step, we have used $\varphi\, \xi^b \pounds_{\varphi\xi}\, h_{ab} = \pounds_{\varphi\xi}\, (\varphi\, \xi^b\, h_{ab}) - h_{ab} \pounds_{\varphi\xi}\, (\varphi\, \xi^b) = 0 - 0 = 0$, in the third step, we have used the definition of the Lie derivative, in the fourth step, we have used $\xi^p\, h_{mp} = 0$ and $\xi^p\, h_{pn}$, and, in the fifth step, we have noted that the first term on the right in brackets gives zero, while, since

$$\Pi_{ab} = h_a{}^m h_b{}^n \nabla_m \xi_n \tag{122}$$

the last two terms each give $\varphi\, \Pi_{ab}$.

Eqn. (121) is analogous to $\dot{x} = (1/2m)p$ in particle mechanics. It expresses the derivative of the "position" (h_{ab}) in terms of the "momentum". Thus, (121) supports the interpretation of Π^{ab} as a "momentum of geometry".

We now obtain two additional equations relating Π^{ab} and h_{ab}. The interpretation follows the derivation. From (122),

$$\begin{aligned} D_a\Pi_{bc} &= h_a{}^m h_b{}^n h_c{}^p \nabla_m [h_n{}^r h_p{}^s \nabla_r \xi_s] \\ &= h_a{}^m h_b{}^n h_c{}^p [(\nabla_m h_n{}^r) h_p{}^s \nabla_r \xi_s \\ &+ h_n{}^r (\nabla_m h_p{}^s) \nabla_r \xi_s + h_n{}^r h_p{}^s \nabla_m \nabla_r \xi_s] \\ &= h_a{}^m h_b{}^n h_c{}^p [(\xi^r \nabla_m \xi_n) h_p{}^s \nabla_r \xi_s \\ &+ h_n{}^r (\xi^s \nabla_m \xi_p) \nabla_r \xi_s + h_n{}^r h_p{}^s \nabla_m \nabla_r \xi_s] \end{aligned} \tag{123}$$

Where, in the second step, we have expanded using the Leibniz rule. Consider the third step above. For the first term, $\nabla_m h_n{}^r = \nabla_m (g_n{}^r + \xi_n \xi^r) = \xi_n \nabla_m \xi^r + \xi^r \nabla_n \xi_m$. But the ξ_n in $\xi_n \nabla_m \xi^r$ gets annihilated by $h_b{}^n$. Thus, we get the first term on the right in the last step above. The second term is identical. Next, note that, since $\xi^s \nabla_m \xi_s = 0$, the second term on the right in (123) vanishes. To get rid of the first term, we antisymmetrize (123) over "a" and "b", so that term vanishes since $\Pi_{[mn]} = 0$. Thus, (123) becomes

$$D_{[a}\Pi_{b]c} = \frac{1}{2} h_a{}^m h_b{}^n h_c{}^p R_{mnpq}\, \xi^q \tag{124}$$

Eqn. (124) is interesting, but it is not quite what we want: we would prefer the Ricci tensor to the Riemann tensor on the right. Hence, we contract (124) over "a" and "c"

$$D^c \Pi_{bc} - D_b \Pi^c{}_c = h_b{}^n R_{nq}\, \xi^q = 8\,\pi\, G\, h_b{}^n \xi^q\, T_{nq} \tag{125}$$

where, in the second step, we have used Einstein's equation.

The second equation we want relates the spatial curvature, $\mathscr{R}_{abcd}$, to other quantities. Let k_c be an arbitrary spatial vector field. Then, by definition of the (spatial) Riemann tensor,

$$\begin{aligned} \frac{1}{2}\mathscr{R}_{abc}{}^d\, k_d &= D_{[a} D_{b]}\, k_c = h_{[a}{}^m h_{b]}{}^n h_c{}^p \nabla_m (h_n{}^r h_p{}^s \nabla_r k_s) \\ &= h_{[a}{}^m h_{b]}{}^n h_c{}^p [(\nabla_m h_n{}^r) h_p{}^s \nabla_r k_s \\ &+ h_n{}^r (\nabla_m h_p{}^s) \nabla_r k_s + h_n{}^r h_p{}^s \nabla_m \nabla_r k_s] \\ &= h_{[a}{}^m h_{b]}{}^n h_c{}^p [(\xi^r \nabla_m \xi_n) h_p{}^s \nabla_r k_s \\ &+ h_n{}^r (\xi^s \nabla_m \xi_p) \nabla_r k_s + h_n{}^r h_p{}^s \nabla_m \nabla_r k_s] \end{aligned} \tag{126}$$

where the steps are exactly those used in (123). As before, $\Pi_{[mn]} = 0$ implies that the first term on the right in (126) vanishes. For the second term on the right, $\xi^s \nabla_r k_s = \nabla_r (\xi^s k_s) - k_s \nabla_r \xi^s = -k^s \nabla_r \xi^s$, since k is spatial. For the third term on the right, use the definition of the (space-time) Riemann tensor. Thus, (126) becomes:

$$\frac{1}{2}\mathscr{R}_{abc}{}^d\, k_d = -\Pi_{c[a}\Pi_{b]d}\, k^d + \frac{1}{2} h_a{}^m h_b{}^n h_c{}^p R_{mnpd}\, k^d \tag{127}$$

Since k_c (spatial) is arbitrary,

$$\mathscr{R}_{abcd} = -2\,\Pi_{c[a}\Pi_{b]d} + h_a{}^m h_b{}^n h_c{}^p h_d{}^q R_{mnpq} \tag{128}$$

Eqn. (128) gives the spatial Riemann tensor, but unfortunately, it involves the entire (space-time) Riemann tensor. To get to the Ricci tensor, we first contract (128) over "a" and "c"

$$\mathscr{R}_{bd} = -\Pi^c{}_c\Pi_{bd} + h_b{}^n h_d{}^q R_{nq} + R_{mbpd}\,\xi^m\xi^p + \Pi^c{}_b\Pi_{cd} \tag{129}$$

We still have space-time Riemann tensor, so we contract again;

$$\begin{aligned}\mathscr{R} &= -\Pi^c{}_c\Pi^d{}_d + \Pi^{cd}\Pi_{cd} + R + 2\,R_{mn}\,\xi^m\xi^n \\ &= -\Pi^c{}_c\Pi^d{}_d + \Pi^{cd}\Pi_{cd} + 8\,\pi\,G\,T_{mn}\,\xi^m\xi^n\end{aligned} \tag{130}$$

where, in the second step, we have used Einstein's equation.

We pause here to discuss the significance of (129) and (130). These two equations involve the data h_{ab} and Π^{ab} (and also the stress-energy), but they do not involve time derivatives. Thus, the data cannot be arbitrary tensor fields (on one of the spacelike three-dimensional surfaces): these fields must satisfy (125) and (130), These are called the constraint equations: they constrain, the acceptable choices for initial data.

We now derive the final equation for the initial-value formulation of general relativity. It is the evolution equation for Π^{ab}:

$$\begin{aligned}\dot\Pi^{ab} &= \pounds_{\varphi\xi}\Pi^{ab} = \pounds_{\varphi\xi}(h^{am}h^{bn}\nabla_m\xi_n) \\ &= 2(\pounds_{\varphi\xi}h^{am})h^{bn}\nabla_m\xi_n + h^{am}h^{bn}\pounds_{\varphi\xi}(\nabla_m\xi_n) \\ &= -4\,\varphi\,\Pi^{am}\Pi^b{}_m + h^{am}h^{bn}[\varphi\,\xi^p\nabla_p\nabla_m\xi_n \\ &+ (\nabla_p\xi_n)\nabla_m(\varphi\,\xi^p) + (\nabla_m\xi_p)\nabla_n(\varphi\,\xi^p)] \\ &= -4\,\varphi\,\Pi^{am}\Pi^b{}_m + h^{am}h^{bn}[\varphi\,\xi^p\nabla_m\nabla_p\xi_n \\ &+ \varphi\,\xi^p R_{pmnq}\,\xi^q + (\nabla_p\xi_n)(\varphi\,\nabla_m\xi^p + \xi^p\nabla_m\varphi) \\ &+ (\nabla_m\xi_p)(\varphi\,\nabla_n\xi^p + \xi^p\nabla_n\varphi)]\end{aligned} \tag{131}$$

where, in the fourth step, we have used $\pounds_{\varphi\xi}\,h^{ab} = -2\,\varphi\,\Pi^{ab}$ (see (121), noting that $h^{ab}h_{cb} = h^a{}_c$), (122), and the expression for the Lie derivative, and in the fifth step, we have used $\nabla_p\nabla_m\xi_n = \nabla_m\nabla_p\xi_n + R_{pmnq}\,\xi^q$. We now do various things to various terms on the right in the last expression in (131). The first term is fine. For the second term, note that

$$\begin{aligned}\xi^p\nabla_m\nabla_p\xi_n &= \nabla_m(\xi^p\nabla_p\xi_n) - (\nabla_m\xi^p)\nabla_p\xi_n \\ &= \nabla_m(-\varphi^{-1}D_n\varphi) - (\nabla_m\xi^p)\nabla_p\xi_n\end{aligned}$$

where we have used (from (118)) $\xi^p\nabla_p\xi_n = -\varphi^{-1}D_n\varphi$. The third term is fine. For the fifth term, use $\xi^p\nabla_p\xi_n = -\varphi^{-1}D_n\varphi$, and, for the seventh term, use $\xi^p\nabla_p\xi_n = 0$. Substituting all this, using (122) repeatedly to replace first derivatives of ξ_a by Π_{ab}, (113) becomes

$$\begin{aligned}\dot\Pi^{ab} &= -3\,\varphi\,\Pi^{am}\Pi^b{}_m + \varphi\,D^b(-\varphi^{-1}D^a\varphi) \\ &= -R^{ambn}\,\xi_m\xi_n\varphi - \varphi^{-1}D^a\varphi\,D^b\varphi\end{aligned} \tag{132}$$

But (132) still contains the space-time Riemann tensor, while we want only Ricci tensor things. We eliminate the Ricci tensor term from (132) by substituting (129) (which gives the offending term in terms of initial data). Making this substitution (and rearranging terms a bit) we obtain, finally,

$$\begin{aligned}\dot{\Pi}^{ab} = &-D^a D^b \varphi - 2\,\varphi\,\Pi^{am}\Pi^b{}_m - \varphi\,\Pi^c{}_c\Pi^{ab} + \varphi\,\mathscr{R}^{ab} \\ &- 8\,\pi\,G\,\varphi\,h^{am}h^{bn}(T_{mn} - \frac{1}{2}T\,g_{mn})\end{aligned} \tag{133}$$

where we have used Einstin's equation.

We have now obtained all the equations for the initial-value formulation of general relativity. (You've probably noticed that we keep doing the same sort of calculating over and over. These techniques are very important, and the derivations are well worth study. Once one gets used to them, the calculations are also very easy.) What remains is to interpret.

Let us consider the source-free case: $T_{ab} = 0$. Then our equations (125), (130), (121), and (133) become, respectively

$$D_b(\Pi^{ab} - \Pi\,h^{ab}) = 0 \tag{134}$$

$$\mathscr{R} - \Pi^{mn}\Pi_{mn} + \Pi^2 = 0 \tag{135}$$

$$\dot{h}_{ab} = 2\,\varphi\,\Pi_{ab} \tag{136}$$

$$\dot{\Pi}^{ab} = -D^a D^b \varphi - 2\,\varphi\,\Pi^{am}\Pi^b{}_m - \varphi\,\Pi\,\Pi^{ab} + \varphi\,\mathscr{R}^{ab} \tag{137}$$

where we have set $\Pi = \Pi^c{}_c$. The first two are called the *constraint equations*, the last two *evolution equations*.

The constraint equations contain neither time-derivatives nor φ. They represent conditions which must be satisfied by our data (h_{ab}, Π^{ab}). The evolution equations give the time derivative of each of h_{ab}, Π^{ab} in terms of the values of these fields. These two equations involve φ, which, physically, represents the "rate at which the evolution proceeds." Thus, φ allows us to evolve (from a given spacelike three-surface) more quickly in some regions than in others. That is, φ represents the fact that our introduction of "time" is unnatural in general relativity.

Suppose we are given a three-dimensional manifold S with a symmetric tensor Π^{ab} on it and a positive-definite metric h_{ab}. This is an initial data set for general relativity. Suppose, further, that these fields satisfy the constraint equations. Choose any field φ on S. Then the evolution equations determine h_{ab} and Π^{ab} an instant dt later. Now pick a φ on this new surface, and again from (136), (137), find h_{ab}, Π^{ab} another instant later. Continuing, one obtains a family of a "stacked up" three-surfaces, with h_{ab} and Π^{ab} on each. In other words, one obtains a spacetime. Our freedom to choose φ at each instant reflects the fact that, from an initial three-surface, there are many ways to draw successive surfaces. From these remarks, one would expect that the constraints are preserved in the evolution, i.e., that the "dots" of (134) and (135) reduce, using (136) and verification is two or three pages of calculations).

Two analogies are represented in the table below:

General Relativity	Particle	Electromagnetism
h_{ab}	x	B position
Π_{ab}	p	E momentum
φ	–	– "evolution rate"
$\nabla_b(\Pi^{ab} - \Pi h^{ab}) = 0$	–	$\nabla \cdot E = 0$ constraint
$\mathscr{R} - \Pi^{mn}\Pi_{mn} + \Pi^2 = 0$	–	$\nabla \cdot B = 0$ constraint
$\dot{h}_{ab} = 2\,\varphi\,\Pi_{ab}$	$\dot{x} = \frac{1}{2m}p$	$\dot{B} = -\nabla \times E$ evolution
$\dot{\Pi}^{ab} = -D^a D^b \varphi - 2\,\varphi\,\Pi^{am}\Pi^b{}_m - \varphi\,\Pi\Pi^{ab} + \varphi\,\mathscr{R}^{ab}$	$\dot{p} = -m\nabla V$	$\dot{E} = \nabla \times B$ evolution

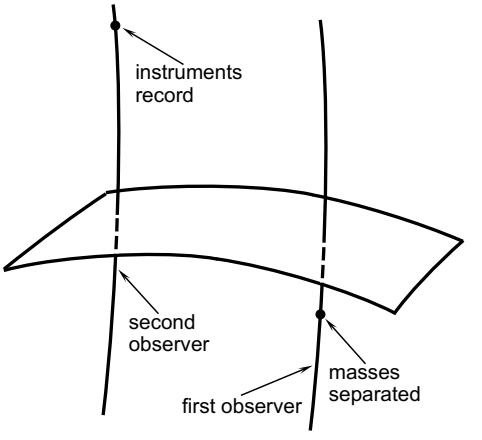
There is nothing in either the case of a particle (in potential V) or in electromagnetism which corresponds to the "evolution rate". The particle and electromagnetic case have a "natural time", and so don't need a special function to represent the rate. There are no constraints for the particle in potential (although there would be if we required that the particle always remain in a certain surface in space). There are constraints in the Maxwell case. Everything has its evolution equations.

In principle, one could imagine discovering solutions of Einstein's equation by finding solutions of (134), (135), and integrating them into the future using (136), (137). In practice, this is not a very fruitful approach.

The existence of an initial-value formulation of general relativity is interesting, in my view, for two reasons: it provides insight into the structure of the theory, and because, in numerous situations in everyday life, one can make an argument or make it clearer by referring to these equations. An example of these reasons follows in the next section.

28. Signal Propagation

What is the speed of propagation of gravitational effect? First note that this is a question which can, in principle, be answered experimentally. Two observers station themselves on separate mountain peaks. One observer holds a rock in each hand at his side, and, at a certain instant, lifts the rock so his arms are outstretched at his sides. From the Newtonian point of view, he produces a mass distribution with a quadruple moment. The observer on the other mountain has sensitive instruments with which he can measure gravitational fields. This second observer records the instant at which his instrument first detect the rock motion. In this way, one could determine a quantity we might call the "speed of propagation of gravitational signals".



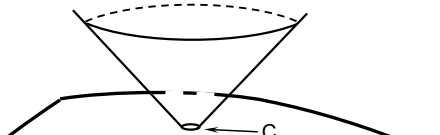
What prediction does general relativity make on this question? Let us first see how, in principle, one might calculate such a thing. One would look around for an exact solution of Einstein's equation in which the stress-energy would reasonably be interpreted as representing "the Earth, with two mountains, two observers, one with a rock in each hand, who separates the rocks, and, with the other observer, some sensitive instruments,..." One would then reproduce, within this space-time, the measurements the observers made when they decided what they meant by the "distance between the mountains", "elapsed time during which the signal traveled", etc. It is clear that this is no way to attack the question.

The initial-value formulation of general relativity offers a comparatively simple way of formulating this question. As an example of an application of this formulation, we now discuss the question of signal propagation.

Note, firstly, that, in our discussion above of a typical experiment, the notion of an initial-value formulation was implicit. The first observer's separating the masses "caused" the second observer's instruments to respond. It was implicitly assumed that what happened in the future (instrument response) was predicated on what had happened earlier (mass separation). It is perhaps not unreasonable, therefore, that the initial-value formulation is appropriate for the discussion.

Let us draw a space-time picture on our experiment. Each observer has a world-line, and we have the events "masses separated" and "instruments respond". We now draw a three-dimensional spacelike surface S in this space-time, where this surface cuts the first observer's world-line just above the event of mass separation. Let us now imagine two situations: the first observer either does or does not separate the masses on cue. Finally, we consider the surface S as a surface on which initial data for our space-time is specified. Now, the question of whether or not the masses were separated determines what the initial data on S is near where the first observer's world-line crosses S. In other words, we can imagine two initial data sets: that when the masses are and are not separated.

Similar experiments reduce to the same question. We are given initial



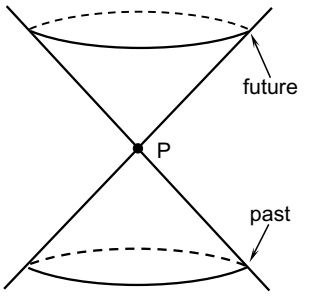
data on a spacelike three-dimensional surface S. We choose a small region C on S. We change the data within C, leaving it alone outside C. In what region of space-time metric changed?

This question is essentially the question of signal propagation. The answer is: the space-time metric is altered (via the evolution equations, within the future light-cone from C. (The proof involves a rather complicated excursion into the properties of hyperbolic partial differential equations.) Thus, gravitational signals propagate at the speed of light (i.e., the light cone "spreads out" at the speed of light, just as gravitational effect do).

In this section, we have shown nothing. We have merely remarked that the initial-value formulation of general relativity leads to a precise and relatively straightforward expression of a question of physical interest. Although the mathematics is still not simple, at least we have the language to ask what we want to ask.

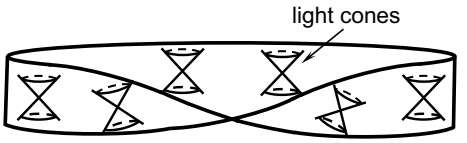
29. Time-Orientation

Let p be an event of the space-time manifold M. The two light-cones at p are qualitatively different physically. One calls one the "future light cone", the other the "past light-cone". One might, therefore, imagine regarding only those space-times as reasonable which admit, globally, such a "past-future distinction". We shall introduce a bit of the mathematics involved in the discussion of such questions.

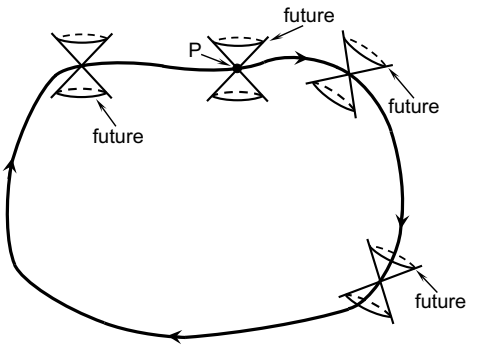


A space-time is said to be *time-orientable* if a continuous choice of a future light-cone can be made globally, throughout space-time. The remarks above suggest that our experiences in our region of the universe provide such a choice in our region of space-time. If, furthermore, we assume that, qualitatively, things are the same everywhere, then we might expect reasonable space-times to be time-orientable.

We first give an example to show that time-orientability is not already a consequence of what a space-time is. It is convenient, to make the drawing of pictures easier, to consider two (rather the four) dimensions. Consider a Mobius strip with light-cones drawn on it as shown. Clearly, a choice of one half of the full light-cone to be called "future" will lead to problem if we try to maintain that choice consistently over the entire space-time. There are never such problems locally.

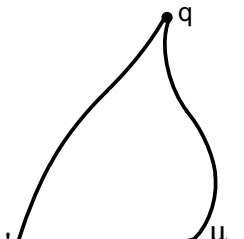


Let us try to formulate the situation more mathematically. Let p be a point of space-time, and let γ be a close curve, beginning and ending at p. At p, make a choice of one of the two light-cones to be called "future". Now carry this choice continuously around the curve γ. It may happen that, on returning to p, (keeping track of the "future light cone" all the way), we find that what we are calling the future light-cone is, as we come back to p (to compare with our original choice) the past light cone. If this were the situation, it would not even be possible to assign "future" for the cones along this single curve – much less over the entire space-time.



We write $L(\gamma) = +1$ if this "light-cone reversal" does not take place on passage around γ, and -1 if it does.

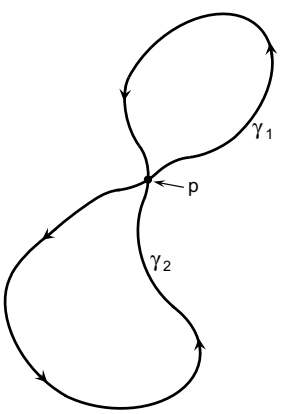
We now claim that "failure of time-orientability takes place either around some closed curve in space-time, or not at all in space-time". More precisely, we claim that a space-time is time-



orientable if and only if $L(\gamma) = +1$ for every closed curve γ. "Proof:" It is clear that, if space-time is time-orintable, then $L(\gamma) = +1$ for every closed curve. Suppose, conversely, that $L(\gamma) = +1$ for every closed curve. Pick any point p in space-time, and there choose one light-cone to be called "future". Let q be any other point of space-time. Draw a curve μ from p to q. Carrying our choice of "future light-cone" continuously along this curve from p, we obtain a choice of a future light-cone at q. Is this choice (at q) independent of the particle curve? Let μ' be another curve from p to q. Then μ and μ' together, define a closed curve beginning and ending at p. But L of this closed curve is $+1$ (assumption). Hence, the designation of future at q is indeed independent of the choice of curve. (This argument can easily be made precise. It is not so easy for many later global arguments.)

Thus, failure of time-orientability is a "curve property". Let's try to formulate that more precisely. Fix a point p of M. Denote by the collection of all closed curves which begin and end at p, where two closed curves such that one can be continuously deformed into the other are regarded as defining the same element of $\Pi_1(M)$. (More precisely: On the collection of all continuous closed curves from p introduce the equivalence relation "are homotopic". Then, $\Pi_1(M)$ is the collection of equivalence classes.) For the cylinder, for example, Π would be just the integers: the integer associated with a closed curve is the number of times it goes around the cylinder before returning to its starting point.

So far, this $\Pi_1(M)$ is just a set. Has it any other structure? We introduce a multiplicative structure. Let, γ_1 and γ_2 be two closed curves, each beginning and ending at p. We introduce a new closed curve, $\gamma_1\gamma_2$ as follows. This $\gamma_1\gamma_2$ is the curve obtained by first following γ_1 back to p, then following γ_2. It is clear that this product is associative: $\gamma_1(\gamma_2\gamma_3) = (\gamma_1\gamma_2)\gamma_3$. The unit curve is the unit which just remains at p – thus of course beginning and ending there. Finally, note that each curve has a "multiplicative inverse", namely the same curve described in the opposite direction.



To summarize, $\Pi_1(M)$ has the structure of a group. It is called the "first homotopy group" of M. (It can actually be defined precisely. We have just given the intuitive flavor of things because the technicalities add little and use time which can perhaps be better spent elsewhere.)

We now make two observations. Firstly, $L(\gamma)$ is invariant under continuous deformation of γ (for L maps into a discrete set, $+1, -1$). Hence, L can be regarded as a mapping from $\Pi_1(M)$ to the set $(+1, -1)$. Secondly, we note that $L(\gamma_1\gamma_2) = L(\gamma_1)\,L(\gamma_2)$. In words, a time-reversing curve composed with a time-preserving curve is time-reversing, etc., for the other two possibilities. In other words, what we have shown is that L is a homomorphism from the group $\Pi_1(M)$ to the (multiplicative) group consisting of $(+1, -1)$. ("Homomorphism" means precisely $L(\gamma_1\gamma_2) = L(\gamma_1)\,L(\gamma_2)$.)

It is clear that the whole story regarding time-orientation (or lack of it) is

summarized by this mapping L. In particular, a space-time M is time-orientable if and only if L maps all of $\Pi_1(M)$ to $+1$.

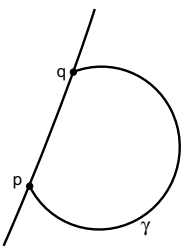
A space-time is said to be *simply connected* if $\Pi_1(M)$ consists of the identity alone (i.e., if every closed curve from p can be continuously deformed to the point p). For example, Minkowski space is simply connected, while a cylinder is not. Clearly, if $\Pi_1(M)$ consists only of the identity e, then $L(e) = +1$ (Proof: $L(ee) = L(e)L(e)$.) Thus, every simply connected space-time is time-orintable.

More, generally, note that the collection of elements of $\Pi_1(M)$ which L sends to $+1$ from a subgroup of $\Pi_1(M)$, and that this subgroup is of order two unless $L[\Pi_1(M)] = +1$. Thus, a space-time is time-orientable if $\Pi_1(M)$ has no subgroup of order two.

The remarks above are only intended to give a taste for how one attacks global questions. One begins (almost always!) with some notion of physical interest. One then tries to express these physical ideas in terms of more or less precise statements about the space-time. Then, one is free to study such statements mathematically, for one has the assurance that results have at least the potential of having physical interest. We shall see a similar approach in each of the next sections.

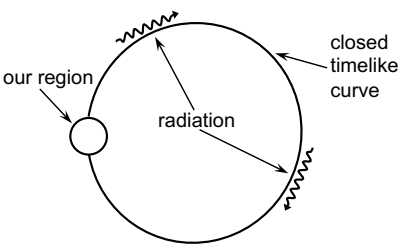
30. Causality Violation

Consider a space-time M, g_{ab} , and suppose there exist within that space-time a closed timelike curve γ. Consider an observer who joins that curve at point p, and leaves it at q. Since the curve continues around from q to p, remaining timelike, our observer has the following opportunity. He could prearrange to have some rocket ship follow the closed curve from q back to p. Inside that ship, he could place a note informing himself to take some action between the events p and q. Thus, the possibility would be available for sending signals into one's own past: one could warn oneself of mistakes one has made in the past, in order to correct them.

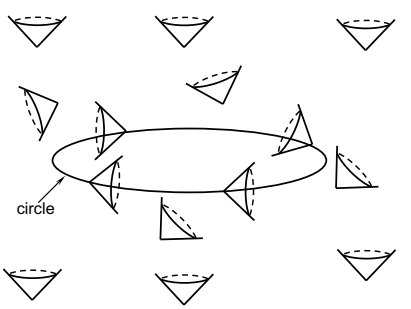


Clearly the presence of closed timelike curves violates our aesthetic principles about how thins are put together. One might be tempted, on that ground alone, to require that no closed timelike curves exist. Only space-time in which this is the case would be regarded as "physically reasonable." But our experiences in physics have taught us to resist such temptations. Almost everything new and broadening in physics appeared, at first sight, to be impossible on aesthetic grounds.

Can more quantitative arguments be made suggesting that closed timelike curves should be ruled out? We make one suggestion in this direction. Suppose our universe had a closed timelike curve passing through our region of space-time. Suppose we, by means of magnets and charges, attempted to obtain a certain electromagnetic field (satisfying Maxwell's equation) in our region of space-time. Since we have closed timelike curves, the effects of such a field would, presumably, propagate around the curve, and re-influence our region of space-time. Thus, one might expect that, in the presence of closed timelike curves, electromagnetic fields must satisfy additional conditions than just Maxwell's equations. Since no such "additional conditions" are observed in Nature, the absence of closed timelike curves is suggested. Problem: Invent a precise, conclusive argument along these lines.



We have seen that the question of the existence of time-orientation is a "curve-question", in the sense that the situation can be adequately described in terms of $\Pi_1(M)$, the first homotopy group of M. Is some similar statement true concerning closed timelike curves? We show, by means of an example, that the answer is no. Our example is a simply connected space-time with closed space-time curves. (We consider three dimensions rather than four, in order to make it possible to draw a figure.) The example is shown in the figure above. In all regions except near a horizontal circle, the light-cones point upward as they usually do. As one approaches the circle, the cones begin to "tip" in the direction



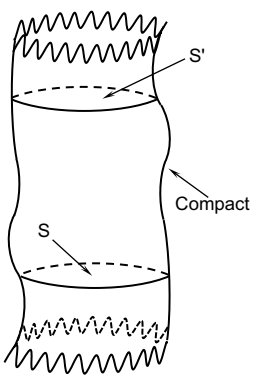
around the circle. On the circle itself, the cones have tipped so far that the circle is our closed timelike curve. But this space is simply connected.

Of course, in the example above, we have merely indicated a qualitative behavior for the light cones. We did not attempt to actually find a solution of Einstein's equation with these features. This is a standard technique in the study of global properties. If something can happen within space-time in general (ignoring Einstein's equation), then it can probably also happen for some solution of the equation. This suppression of the Einstein equation, of course makes things easier to think about. The phenomenon above, for example, actually occurs for certain solution, e.g., the Gödel solution.

31. An Implication of Absence of Causality Violation

What sort of statements about the global structure of our universe follow if one assumes absence of closed timelike curves? We give an example of one such statement in the present section.

Consider a space-time manifold M with metric g_{ab}. Let us suppose that there are spacelike surfaces S and S' in our space-time. (Physically, each represents *all of space at one instant of time*.) We suppose, furthermore, that the region of space-time between S and S' is compact. (A region of space-time is said to be compact if every sequence of points in that region, $P_1, P_2, \ldots$ has a sub-sequence, e.g., $P_7, P_{19}, P_{237}, \ldots$ which converges to some point of the region. A compact region of space-time is closed, and, intuitively, it does not "go off to infinity".)

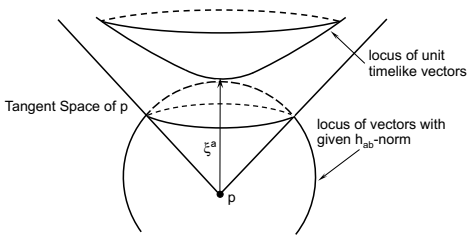


The closed (i.e., positive spatial curvature) Friedmann models provide an example of a space-time satisfying these conditions. For S and S', take the spatial sections. Then, since these sections are three-dimensional spheres, the region between two such sections is compact. Physically, we are considering cosmological models which are "spatially closed".

We shall show below that, if such a space-time is time orientable, and has no closed timelike curves, then S and S' are diffeomorphic (i.e., identical as manifolds). Physically, this means that time orientability and causality nonviolation (physically reasonable conditions) imply that the "topology of space" cannot change from one epoch to another in a closed universe. This result, in fact, does not require Einstein's equation: it just depends on a few geometrical properties of light cones.

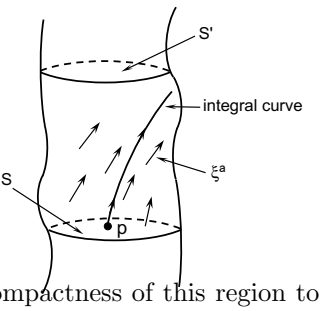
Introduce an arbitrary positive-definite metric h_{ab} on M (Such always exists. This is a moderately subtle theorem about manifolds. The proof proceeds along the following lines: existence of g_{ab} on M implies that there is a derivative operator on M; existence of a derivative operator implies that M is paracompact; that M is paracompact implies existence of positive-definite metric.)

The next step in the construction is to obtain an everywhere timelike vector field on M. Fix a point p of M. Among all vectors ξ^a at p which are unit, future-directed, and timelike (with respect to g_{ab}), consider that one which has the smallest norm with respect to h_{ab}. This is our ξ^a. (Geometrically, the unit, future-directed timelike vectors at p describe a hyperbola in the tangent space. The vectors of a given norm with respect to h_{ab} describe a sphere. We adjust the radius of this sphere so that it intersects the hyperboloid in just one point.) Repeating at each point of our space-time, we obtain our desired timelike vector field on M. (This vector



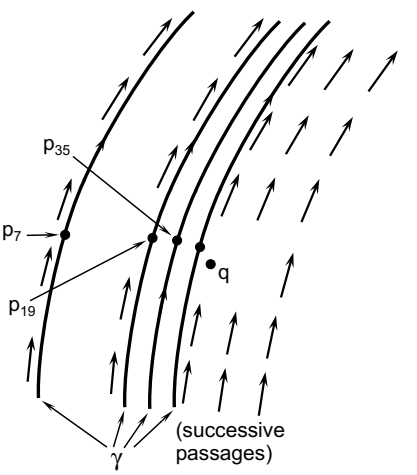
field has no direct physical significance. It merely provides a convenient way of keeping track of what the light cones are doing.)

Next, choose any point p of S, and draw the integral curve γ which begins at p. What can happen to this curve? It could return to S, remain within the region between S and S', or go to S'. The idea is to show that the first two possibilities cannot occur.



The curve γ cannot return to S, for our timelike vector field ξ^a points from S into the region between S and S'. We use compactness of this region to show that this possibility violates our assumption of no closed timelike curves. Choose a sequence of points $p_1, p_2, \ldots$ along the curve γ such that every point of γ lies between two points of this sequence. Here is a sequence on a compact region, so some subsequence approaches a point q of the region. Thus, the situation is that pictured on the right. Our curve continually sweeps by the point q, coming closer and closer, so a subsequence of our sequence approaches q. But now it is clear that we have a closed timelike curve in our space-time: merely distort γ on two of its successive passages near q to obtain a closed timelike curve, beginning and ending at q.

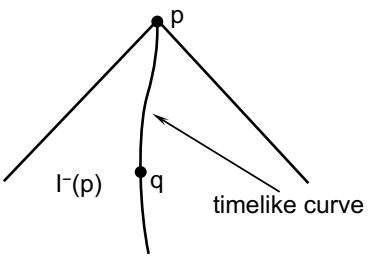
Thus, our curve cannot remain indefinitely in the region between S and S' without there existing closed timelike curves. Hence, γ goes to S'. But this is true for the curve beginning at each point of p and S. Thus, we obtain a mapping from S to S'. Repeating the same argument, from S' to S, we obtain an inverse mapping. This shows that S and S' are equivalent as manifolds. (More correctly, one would have to show that this mapping is smooth, with smooth inverse, to establish that it is a diffeomorphism.)



The argument above is intended merely as one example of how one proceeds to extract global information from an assumption of no casual anomalies. We repeat the conclusion in intuitive terms: "In a spatially closed, time-orientable, causally reasonable space-time, the topology of space cannot change from one epoch to another."

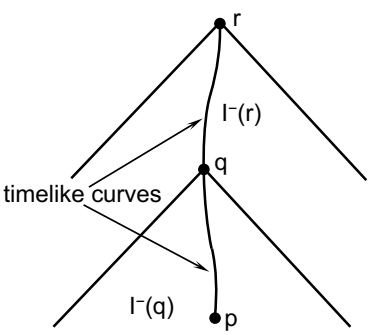
32. The Domain of Influence

In this section, we study the mathematical description of the question "Does what happen at event p influence what happens at event q?", i.e.," Can a signal be sent from event p to event q?" This is clearly a fundamental relation between events. Again, our purpose is merely to give the flavor of how one describes things and what conclusion can be drawn.

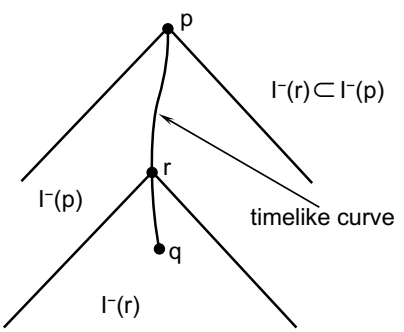


Let M be a time-orientable space-time without closed timelike curves. Let p be a point of M. We denote by $I^-(p)$ the collection of all points q of the space-time which can be joined to p by a future-directed timelike curve. (Thus, $I^-(p)$ is what we have called the "interior of the past light-cone of p".) This $I^-(p)$ will be called the *past* of p, or the *past domain of influence* of p. (The latter term is perhaps unfortunate. $I^-(p)$ does not include its boundary, whereas points on the boundary can, in general, also influence p.) Physically, $I^-(p)$ represents the collection of all events of space-time which can affect what happens at p. Similarly, $I^+(p)$, the *future* of p, or *future domain of influence* of p, is the collection of all points which can be joined to p by a past-directed timelike curve.

It is clear that I^- and I^+ provide a concise formulation of the possibilities for events influencing other events. What remains is to describe some properties of I^+ and I^-.



We first remark that p is in $I^-(q)$ if and only if q is in $I^+(p)$. (That is, a point p can be joined to q by a future-directed timelike curve if and only if q can be joined to p by a past-directed timelike curve.) Furthermore, if p is in $I^-(q)$, and q is in $I^-(r)$, then p is in $I^-(r)$. (Proof: Join the future-directed timelike curve from p to q to the future-directed timelike curve from q to r. The result is a future-directed timelike curve from p to r.)

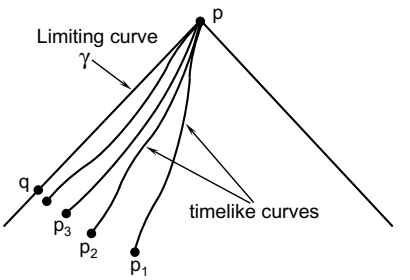


We remark that $I^-(p)$ is always an open subset of space-time. This fact is suggested by the following argument. Firstly, in Minkowski space, $I^-(p)$ is open. Hence, if p is a point of a general space-time, one would expect that $I^-(p)$ is open in some neighborhood of p. (For locally, the metric is essentially that of Minkowski space.) (Explicitly, one could introduce a chart in a neighborhood of p, explicitly write $I^-(p)$ in terms of coordinates, and verify that it is open.) Thus, what remains is to show that this "local openness of $I^-(p)$ implies global openness". Let q be a point of our space-time in $I^-(p)$, so we have a future-directed timelike curve from q to p.

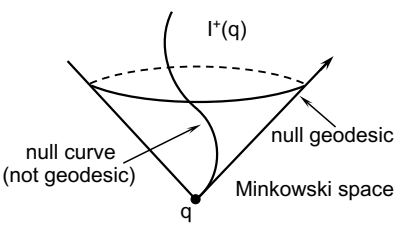
Choose a point r on this curve near q. Then, by construction, r is in $I^-(p)$, and q is in $I^-(r)$. But $I^-(r)$ is open near r, so q is in this open set. But $I^-(r)$ is a subset of $I^-(p)$. Thus, we have obtained an open subset of $I^-(p)$ which contains q. Since q was an arbitrary point of $I^-(p)$, $I^-(p)$ is open.

The boundary of $I^-(p)$ is, roughly speaking, the "past null cone" of p. One might expect, in particular, that it is a null surface. In fact, this is true in a sense: $I^-(p)$ contains null geodesics. To prove this, let q be a point of the boundary of $I^-(p)$. Choose a sequence of points $p_1, p_2, \ldots$ of $I^-(p)$ which approach q. Since each of these points are in $I^-(p)$, each can be joined to p by a future-directed timelike curve.

Let γ be the limit of this sequence of curves. Now, γ cannot be a timelike curve, for, if it were, then q would lie in $I^-(p)$, which, since $I^-(p)$ is open, would contradict the fact that q is on the boundary of $I^-(p)$. Thus, γ must be a null curve. Furthermore, γ cannot enter $I^-(p)$, for this, again would cause q to be in $I^-(p)$. Hence, γ must lie on the boundary of $I^-(p)$. Next, note that γ cannot enter the future of q, for, if it did so, one of the timelike curves from $p_1, p_2, \ldots$ must enter the future of q (since $I^+(q)$ is open, and the sequence of timelike curves approaches γ). This implies that γ is a null geodesic. (A null curve from q in Minkowski space which is not a null geodesic enters the future of q. Thus, at any point at which γ is not a null geodesic, there γ would enter the future of q, by a local argument.)

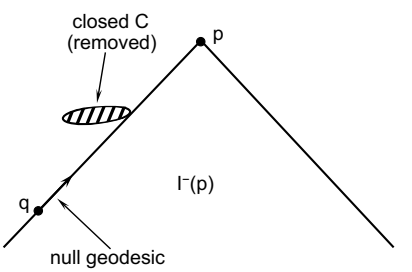


To summarize, if q is in the boundary of $I^-(p)$, then there is a future-directed null geodesic from q which lies in the boundary of $I^-(p)$. In this sense, then the boundary of $I^-(p)$ is a "null cone".



What can happens to this null geodesic? It may, as in Minkowski space, remain in the boundary of $I^-(p)$ until it reaches p. It is also possible that it never reach p. To construct an example of this behavior, consider Minkowski space with the closed region C removed from the manifold. In this example, the null geodesic from q meets C, and, since C has been removed, does not continue beyond this meeting. Thus, the null geodesic never reaches p.

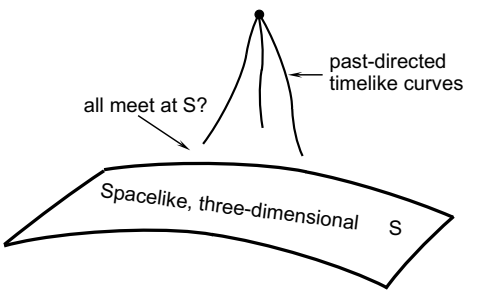
To summarize, the structure suggested by the question "Can event p influence event q" is perhaps more fundamental than the manifold and metric structure which forms the basis for general relativity. This fact suggests that this causal structure be studied in its own right. The mathematical description is in terms of $I^+(p)$ and $I^-(p)$, the future and past. We have defined these sets in a (time-orientable, no closed timelike curves) space-time, and obtained a few of their properties.



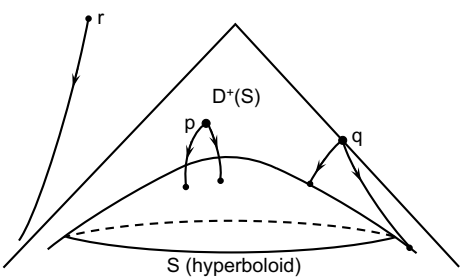
33. The Domain of Dependence

In Sect. 32, we discussed the question "Can event p influence event q?" We now consider the question "Does what happens on a surface S completely determines what happens at event q"? This leads to the notion of the domain of dependence. Again, our purpose is merely to introduce the definition and obtain a few properties. We shall see that the domain of dependence is, in a sense, dual to the domain of influence.

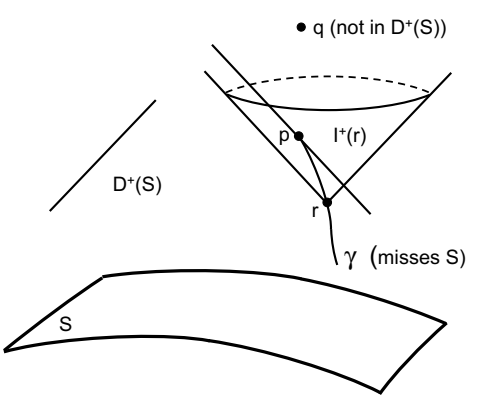
Let M be a time-orientable space-time without closed timelike curves. Let S be a closed, three-dimensional, spacelike surface in M. Let p be a point to the future of S. Under what conditions would we expect what happens at p to be completely determined by information on S? Recall that signals in general relativity propagate along timelike or null curves. Thus, we define the *future domain of dependence* of S, $D^+(S)$ as the collection of all points p of space-time such that every past-directed timelike curve from p meets S. (The definition above would make more sense physically if we replaced "timelike" by "timelike or null". It's only a question of whether or not $D^+(S)$ contains its boundary. For mathematical reasons, the above definition is more convenient.) Similarly, we define $D^-(S)$. (S will be included in $D^+(S)$ and $D^-(S)$.)



We give two examples of the domain of dependence. Let S be the surface $t = 0$ (in the usual coordinates) in Minkowski space. Then $D^+(S)$ is the region $t > 0$ in Minkowski space. For a less trivial example, let S be a hyperboloid in Minkowski space. Then $D^+(S)$ consists of all points between (and including) the hyperboloid and the light cone. Points p and q, for example, are in $D^+(S)$, for every past-directed timelike curve from either point reaches S. Point r is not in $D^+(S)$, because of past-directed timelike curve shown.



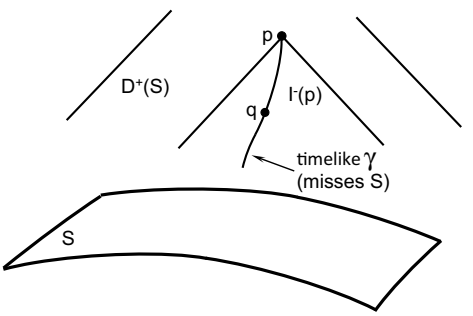
We now derive a few elementary properties of $D^+(S)$. The first is that $D^+(S)$ is closed. Let p be a point on the boundary of $D^+(S)$. Suppose that p were not in $D^+(S)$. Then there would be a past-directed timelike curve γ from p which fails to meet S. Choose a point r on this curve slightly to the past of p. Then r is is not in $D^+(S)$ (for the rest of γ is a past-directed timelike curve from r which fails to meet S.) Hence, any point in the future of r is not in $D^+(S)$ (for, if q is in $I^+(r)$, then there is a past-directed timelike curve from q to r. Continue this curve along γ, so it misses S. We thus



obtain a past-directed timelike curve from q which fails to meet S, so q is not in $D^+(S)$.) Now, $I^+(r)$ is open, contains p, and fails to meet $D^+(S)$. This contradicts our original placement of p on the boundary of $D^+(S)$. Hence, $D^+(S)$ is closed.

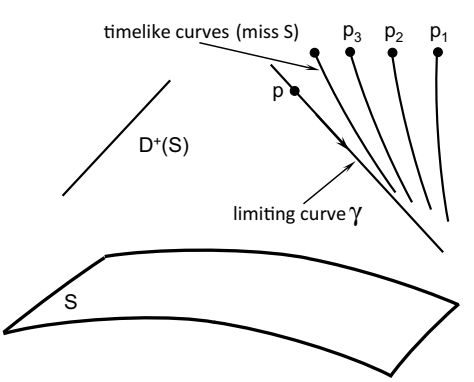
We next remark that, if p is in $D^+(S)$, and q is in $I^-(p)$, and in $I^+(s)$ for some point s of S, then q is in $D^+(S)$. Proof: suppose not. Then there is a past-directed timelike curve from q which misses S. Since q is in $I^-(p)$, there is a past-directed timelike curve from p to q. Joining these, we obtain a past-directed timelike curve from p which misses S. This contradicts our original placement of p in $D^+(S)$.

If follows immediately that, if p is in the boundary of $D^+(S)$, then $I^+(p)$ does not intersect $D^+(S)$. (Proof: If q is in $I^+(p)$ and $D^+(S)$, then p is in $I^-(q)$. By the remark above, all of $I^-(q)$ to the future of S is in $D^+(S)$. Hence, we have an open set in spacetime which is in $D^+(S)$ and which includes p. This contradicts the placement of p on the boundary of $D^+(S)$.)

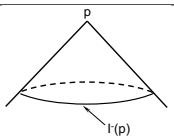
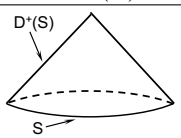


As a final property of $D^+(S)$, we establish the following: If p is in the boundary of $D^+(S)$, then there is a past-directed null geodesic γ from p which remains in the boundary of $D^+(S)$. The proof is analogous to the proof that the boundary of the past of a point contains null geodesics. Choose a sequence of points, $p_1, p_2, \ldots$, which approach p, and which are not in $D^+(S)$. Then from each there is a past-directed timelike curve which fails to meet S. Take the accumulation curve of this sequence, and call it γ. (There is a technical argument that such an accumulation curve must exist. The key step is to fix a compact neighborhood of p, locate the points at which these curves first leave that neighborhood, and find an accumulation point of that sequence.) Now, this γ must be timelike or null (since it is a limit of timelike curves). This γ cannot enter the interior of $D^+(S)$, for if it did, some timelike curve from some of $p_1, p_2, \ldots$ would enter $D^+(S)$, from which such a curve would necessarily meet S. Hence, γ cannot enter $I^-(p)$ to the future of S, for as we have seen, this open region is in $D^+(S)$. Thus, γ is a timelike or null curve which remains on the boundary of $I^-(p)$: hence γ is a null geodesic. Since γ never enters the interior of $D^+(S)$, and since γ remains of the boundary of $I^-(p)$ (which is in $D^+(S)$), γ must remain on the boundary of $D^+(S)$.

We conclude that, through every point of the boundary of $D^+(S)$, there passes a past-directed null geodesic which remains on the boundary of $D^+(S)$. See the first figure on page 162.



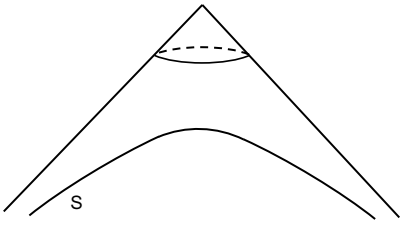
The domain of influence and the domain of dependence have many properties in common. This analogy is represented by the table below.

Concept	Domain of Influence	Domain of Dependence
Depends on	Point p $I^-(p)$	Spacelike Three-Dimensional Surface S $D^+(S)$
Figure		
Topology	Open	Closed
Boundary Point q	$I^-(q) \subset I^-(p)$ $I^+(q)$ not intersect $I^-(p)$	$I^-(q)$ (in future of S) $\subset D^+(S)$ $I^+(q)$ not intersect $D^+(S)$
Null Geodesics Through Boundary Point q	Future-directed, remains in boundary $I^-(p)$	Past-directed, remains in boundary $D^+(S)$

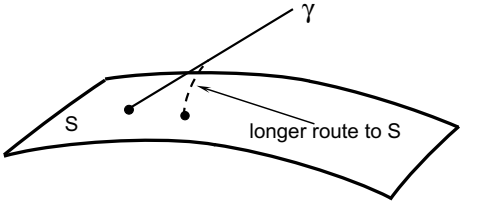
34. Singularity Theorems

In the last ten years or so, a number of results have been obtained to the effect that solutions of Einstein's equation have a tendency to become singular. We have already seen an example of singular behavior in space-time: in the Friedman solutions. There are rather analogous to the singular behavior one would see in Newtonian theory for an initially static ball of dust which is spherically symmetric – and which is allowed to collapse. All the dust particles fall toward the center, and there the density eventually becomes infinite. Why does nobody get excited about this Newtonian phenomenon? Because it is not "generic". If, for example, one endows our ball of dust with a small bit of angular momentum initially, then the ball will collapse a certain amount, but eventually the angular momentum will take over, and the ball will reexpand. The singularity in this Newtonian example arises only because of the special symmetry. Is a similar remark applicable to the singularity in cosmological (i.e., Friedman) solutions? This was for a long time a controversial question in general relativity. As we shall see in this section, the answer is no.

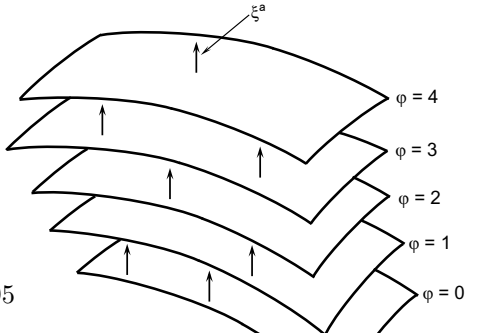
Let (M, g_{ab}), be a space-time, and let S be a three-dimensional, spacelike surface in M. Let p be a point of M to the future of S. We shall be concerned with the following question: Does there exist a timelike curve from p to S whose length in maximal? In order to have an example before us, consider a hyperboloid S in Minkowski space. As we shall see shortly, it is the points between S and the light-cone (e.g., p in the figure, but not q) for which there does exist a longest timelike curve to S.



We return now to the general case. Suppose, somehow, that we did manage to find a timelike curve γ from p to S whose length is maximal. Then γ must be a geodesic (for otherwise we could lengthen γ by straightening it out; recall that a timelike curve which "wiggles" gets near the light cone, and so tends to have a smaller length), and must intersect S orthogonally (for otherwise we could lengthen by moving slightly into point of intersection with S; again, this is a statement which is obvious in Euclidean space, except that "lenghten" replaces "shorten"). Thus, a longest timelike curve from p to S – if one exists – is a geodesic which meets S orthogonally. We are therefore led to study these normal geodesics.



We introduce the function φ (defined, at least, near S) such that $\varphi(p)$ is the distance from S along the normal geodesic. Set $\xi_a = \nabla_a \varphi$. (This vector field, although it will play an important role, has little direct physical significance.) Then, since φ measures distance, we have $\xi^a \xi_a = -1$. Note, furthermore, that $\nabla_{[a}\xi_{b]} = \nabla_{[a}\nabla_{b]}\varphi = 0$.



That is, $\nabla_a\xi_b$ is symmetric. Hence, $\xi^b\nabla_b\xi_c = \xi^b\nabla_c\xi_b = \frac{1}{2}\nabla_c(\xi^b\xi_b) = \frac{1}{2}\nabla_c(-1) = 0$. Therefore, ξ^a is tangent to geodesics. Since, furthermore, ξ^a is normal to S (since $\varphi = 0$ on S), the integral curves of ξ^a are precisely the normal geodesics from S – the candidates for the longest timelike curves from points of M to S.

An important little calculation underlines all the singularity theorems. Set $c = -\nabla_a\xi^a$, the convergence of ξ^a. We are concerned with the derivative of c along our geodesics:

$$\begin{aligned}\xi^m\nabla_m c &= -\xi^m\nabla_m\nabla_a\xi^a = -\xi^m\nabla_a\nabla_m\xi^a - \xi^m R_{ma}{}^a{}_b\,\xi^b \\ &= -\nabla_a(\xi^m\nabla_m\xi^a) + (\nabla_a\xi^m)(\nabla_m\xi^a) + \xi^m\xi^b R_{mb} \\ &= (\nabla^a\xi^b)(\nabla_b\xi_a) + \xi^a\xi^b R_{ab}\end{aligned} \tag{138}$$

where, in the second step, we have used the definition of the Riemann tensor; in the third step, we have differentiated by parts; and, in the fourth step, we have used the definition of the Ricci tensor and the fact that $\xi^m\nabla_m\xi^a = 0$. We next modify each term on the right in (138). For the first term, note that $\nabla_a\xi_b$ is a symmetric, spatial (with respect to ξ^a) tensor.

Write $\nabla_a\xi_b = -(1/3)c\,h_{ab} + (\nabla_a\xi_b + (1/3)c\,h_{ab})$, where $h_{ab} = g_{ab} + \xi_a\xi_b$. That is, we decompose $\nabla_a\xi_b$ into its "trace part" plus its "trace-free part". Then

$$\begin{aligned}(\nabla_a\xi_b)(\nabla^a\xi^b) &= [-\frac{1}{3}c\,h_{ab} + (\nabla_a\xi_b + \frac{1}{3}c\,h_{ab})][-\frac{1}{3}c\,h^{ab} + (\nabla^a\xi^b + \frac{1}{3}c\,h^{ab})] \\ &= \frac{1}{3}c^2 + (\nabla_a\xi_b + \frac{1}{3}c\,h_{ab})(\nabla^a\xi^b + \frac{1}{3}c\,h^{ab})\end{aligned} \tag{139}$$

But the second term on the right in (139) is non-negative. (Proof: The trace of the square of a symmetric 3×3 matrix is non-negative.) Hence,

$$(\nabla_a\xi_b)(\nabla^a\xi^b) \geq \frac{1}{3}c^2 \tag{140}$$

To treat the second term on the right in (138), we use Einstein's equation:

$$\xi^a\xi^b R_{ab} = 8\,\pi\,G(T_{ab} - \frac{1}{2}T g_{ab})\xi^a\xi^b \tag{141}$$

In the Newtonian limit, the right side in (141) is essentially the mass density (as seen by an observer with four-velocity ξ^a). Hence, this is positive in this limit. We now assume that the matter in our space-time is such that the right side of (141) is non-negative. This is called the *energy condition*. The energy condition is reasonable physically, firstly, because no matter has ever been observed which violates this condition, and secondly, because no theoretical models for matter have ever been constructed violating the energy condition. For a perfect fluid, for example the energy condition requires $\rho + p$ and $\rho + 3p$ be non-negative. For an electromagnetic field, the right side of (141) is $4\,\pi\,G(E^2 + B^2)$ (using E and B relative to ξ^a), and hence non-negative.

Assuming that the energy condition is satisfied, (138) becomes,

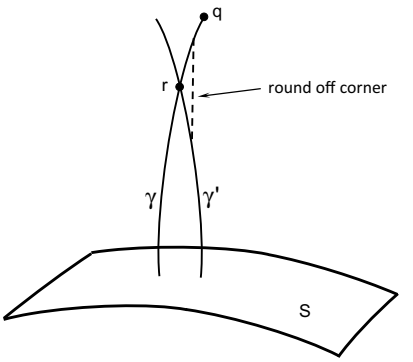
$$\xi^m\nabla_m c \geq \frac{1}{3}c^2 \tag{142}$$

But (142) can be solved (dividing by c^2). Let τ be proper length along our geodesic, with $\tau = 0$ on S. Then the solution, $c(\tau)$, of (142) is (where $c_0 = c(0)$, the value on S)

$$c(\tau) \geq \frac{3}{3/c_0 - \tau} \tag{143}$$

Physically, (142) says that there is an irreversible tendency for our geodesics to converge – provided the energy condition is satisfied. The energy condition thus ensures that the matter acts "attractively", i.e., so as to pull the geodesics closer and closer together. This motion is seen even more explicitly in (143). We see that c becomes infinite at least by $\tau = 3/c_0$.

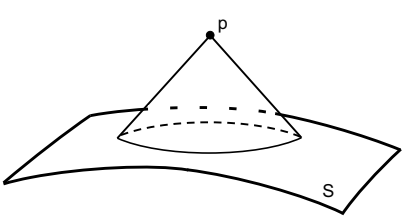
What does it mean geometrically geometrically when c becomes infinite? How can the convergence become infinite? It means that our geodesics have begun to cross (i.e., form a caustic). Of course, this crossing of timelike geodesics does not by itself indicate any singular behavior in our space-time. Geodesics, for example, can cross in Minkowski space. In the figure of page 161, for example, the normal geodesics from S all meet at p – but p is a perfectly regular point of space-time. We are interested in this crossing of geodesics for a more subtle reason: because crossing says something about the possibility of a geodesic's being the longest curve to S. Suppose two nearby normal geodesics from S cross. Then γ cannot be the longest timelike curve from q to S, for we can find a longer curve by "rounding off the corner" at the crossing point r, and then following γ' down to S.



We stop to summarize what we have so far. We have seen what, if $c \geq c_0 >$ (c_0 a constant) on S, then every normal geodesic from S meets another geodesic at least by the time it has gone a distance $3/c_0$. (Here, the energy condition was required.) We have seen, furthermore, that, if one normal geodesic crosses another normal geodesic, then the first geodesic cannot be the longest timelike curve from a point beyond the crossing point to S. Finally, we have seen that, if there is a longest timelike curve from a point p to S, then that curve must be a geodesic which meets S orthogonally. Taken together, these statements imply that, *if we go farther than* $3/c_0$ *from* S *along any timelike curve, to some point* p*, then there is no longest timelike curve from* p *to* s*.* (For, if there were such a longest curve, it would be anormal geodesic with length greater than $3/c_0$. But such cannot be the longest.)

The underline statement above is all we need to complete the proof. It states that, sufficiently far away from S, the points of space-time have the property that they cannot be joined to S by a longest timelike curve.

We need one final assumption: that $D^+(S)$ is the entire future of S. That is to say, we now assume that our space-time is such that it is completely determined by what happens on S. (Such an S is called a *Cauchy surface.*) The assumption that S is a Cauchy surface implies that, from any p in $D^+(S)$, there is a longest timelike curve to S. This, again is a rather complicated technical result, but at least we can indicate why it is true. If p is in $D^+(S)$, then every past-directed timelike curve from p meets S. It can be shown from this that the closure of the union of these curves is compact. (Intuitively, this region cannot



"go off to infinity," or else some timelike curve from p would "go off to infinity" missing S.) One then shows that the collection of all timelike or null curves from p to S is compact (in a suitable topology), and that length is a continuous function on this compact space of curves. Hence, "length" achieves its maximum.

We are now essentially done with the proof. If we go a distance $3/c_0$ from S along any timelike curve, we reach a point from which there is no longest timelike curve to S. But, if we assume that S is a Cauchy surface, then from every point, there is a longest timelike curve to S. But, if we assume that S is a Cauchy surface, then, from every point, there is a longest timelike curve to S. How can these statements be consistent? Only if one simply cannot go a distance more than $3/c_0$ from S along any timelike curve. Thus:

Let a space-time satisfy Einstein's equation, with a stress-energy satisfying the energy condition. Let that space-time have a Cauchy surface S on which $c \geq c_0$. where c_0 is a positive constant. Then no future-directed timelike curve from S has length greater than $3/c_0$.

There are two issues about this result which require discussion: What has it to do with our universe? and what has it to do with singularities?

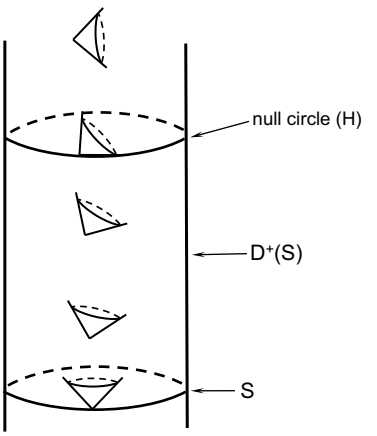
We have remarked that the Friedmann solutions appear to represent a good approximation to our universe. Let S be one of the spacelike section in such a model. Then S is indeed a Cauchy surface, but the value of c on this surface is $-\alpha$ (where α is the Hubble function). Thus, c is not positive in the present epoch. The idea is to apply the result above, but with the roles of past and future interchanged. Thus, the convergence into the past is $\alpha > 0$, and all the conditions are satisfied. Thus, the result above states that timelike curves into the past cannot have length than $3/c_0$, and, indeed, in our Friedmann solutions, timelike curves into the past ran into the singularity, when ρ becomes infinite. The advantage of the result on the previous page over the explicit remark about the Friedmann solutions is that the result makes no assumptions about symmetric or other details of the solution – it is applicable to all solutions which display certain broad features. Thus, if we perturb (slightly) a Friedmann model (our universe, of course, would be a perturbed Friedmann solution, e.g., because isotropy does not hold exactly) we still must obtain the result that timelike curves into the past are no longer than a certain amount.

What has the result to do with singularities? The problem here is that we have required from the beginning that space-time be a smooth manifold with a smooth metric , and other smooth tensor fields. Thus, there is no way to represent singularities directly on a space-time manifold. (To give up smoothness leads to serious difficulties of principle.) If there were "actual singular points" on the space-time manifold, then such points would have to be removed to obtain the required smoothness. How would this "removal" be detected, given only the smooth space-time (after removal of all "singular points")? It could be detected by noting that certain timelike curves (namely, those which would have passed through the singular region) have only finite length in space-time. Thus, to describe singular behavior in space-time, one looks, not for certain quantities becoming infinite, but rather for timelike curves of finite length. Physically, an observer who follows such a curve has a finite total lifespan – "after that" he is no longer representable by a point on the space-time manifold. From the point of view of this observer, something singular is certainly going on.

The underline statement on the preceding page is rather weak. The problem lies with the assumption that S is a Cauchy surface. Suppose we are given initial data on S, and that we evolve this data to obtain a solution of Einstein's equation. It may just turn out that S will not be a Cauchy surface – that our universe is such that it is not completely determined by what goes on on a certain spacelike three-dimensional surface. Consider, for example, the hyperboloid on page 165. The

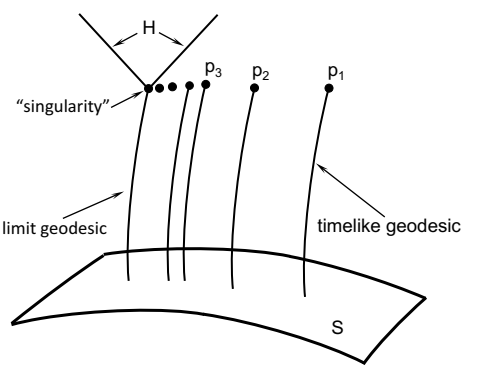
future domain of dependence for this S is the region between S and the light cone. It is not a Cauchy surface. Quite generally, singularity theorems which assume a Cauchy surface are rather weaker than those which do not. We now indicate how our theorem can be modified to eliminate the Cauchy surface assumption.

Assume that S is compact. Denote by H the boundary of the domain of dependence of S (S itself excluded).The crucial question is; Is this H also compact? Either answer leads to a difficulty.



Suppose first that H is compact. A two-dimensional space of this type is shown in the figure. The space-time is a cylinder, with the light-cones invariant under rotations about the axis. Near the bottom of the cylinder, the light-cones point up the cylinder. Further up, the light cones begin to "tip", so that, at a certain point, a circular cross-section of the cylinder is a null curve. After that, the cones continue to tip, giving closed timelike curves. the domain of dependence of S is the region between S and the null circle, while the null circle is H. This H is compact. The difficulty in this example is that we have closed timelike curves. We now show that a similar phenomenon occurs in the general case. We have seen in Sect. 33 that there are null geodesics which lie in H. Since H is compact, such a null geodesic must wander around in H until it eventually comes back arbitrarily close to itself. We have almost-closed null curves – a type of causality violation as serious, for all practical purposes as closed timelike curves. (There is an exact solution of Einstein's equation – called Taub-NUT space – in which precisely this phenomenon occurs.)

Now suppose that H is not compact. Then we can find a sequence of points, $p_1, p_2 \ldots$ in $D^+(S)$ which do not have a point of accumulation. Since each point is in $D^+(S)$ – where our earlier argument applies – from each point there is a past-directed timelike geodesic which meets S orthogonally, and which has length no greater than $3/c_0$. Take the limit of this sequence of geodesics. The result in another timelike geodesic. Clearly, the length of this geodesic cannot be greater than $3/c_0$, for, if it were, it would provide a point of accumulation for $p_1, p_2 \ldots$. Thus, we obtain a timelike geodesic from S of length no greater than $3/c_0$.



To summarize, we have

Let a space-time satisfy Einstein's equation, with a stress-energy satisfying the energy condition. Let that space-time have a three-dimensional spacelike surface S on which $c \geq c_0$, where c_0 is a positive constant. Let there be no almost closed null curves. Then some future-directed timelike geodesic from S has length no greater than $3/c_0$.

The first result requires a Cauchy surface, while the second only requires a compact spacelike surface. (That is to say, the second is applicable only to "closed

universes.") The second result requires a causality condition. Finally, the first result gives a finite length for every timelike curve from S, while the second gives a finite length for just one timelike geodesic from S.

35. Conformal Transformation

In this section, we introduce a mathematical tool which finds numerous applications in general relativity.

Let (M, g_{ab}) be a space-time. A *conformal transformation* on g_{ab} consists of replacing g_{ab} by the metric $\tilde{g}_{ab} = \Omega^2 g_{ab}$ where Ω is a positive scalar field on M. We call Ω the *conformal factor*. Geometrically, a conformal transformation amounts to "rescaling distances without changing angles". it is normally the case in practice that the new metric $\tilde{g}_{ab}$ has no direct physical significance. (It is, in fact, sometimes called the "unphysical metric".) A conformal transformation in merely a mathematical convenience.

Let, $T^{a\cdots c}{}_{b\cdots d}$ be some tensor field on M. Of course, the notion of a tensor field makes no direct reference to a metric. Hence, we still have this tensor field $T^{a\cdots c}{}_{b\cdots d}$ on M after some conformal transformation. It is convenient, however, to allow ourselves the following option. For a given conformal transformation $\tilde{g}_{ab} = \Omega^2 g_{ab}$, we set $\tilde{T}^{a\cdots c}{}_{b\cdots d} = \Omega^n T^{a\cdots c}{}_{b\cdots d}$ where n is some real number (in practice, usually an integer). That is, we allow other tensor fields to be "rescaled" along with the metric. The number $m = n-$ (number of covariant indices on T) + (number of contravariant indices on T) is called the *dimension* of $T^{a\cdots c}{}_{b\cdots d}$. (Often, the term dimension refers to $\sec^m$ rather than the number m of itself.) For example, the metric g_{ab}, considered as just another tensor field, has dimension $0\ (= 2 - 2 + 0)$.

Suppose we have selected a dimension for $T^{a\cdots c}{}_{b\cdots d}$ and $S^{p\cdots q}{}_{r\cdots s}$, (i.e., suppose we have decided what $\tilde{T}^{a\cdots c}{}_{b\cdots d}$ and $\tilde{S}^{p\cdots q}{}_{r\cdots s}$ will be under a conformal transformation). What dimension should one choose for the tensor field $T^{a\cdots c}{}_{b\cdots d}\, S^{p\cdots q}{}_{r\cdots s}$. The natural choice is to assign to this outer product a dimension equal to the sum of the dimensions of $T^{a\cdots c}{}_{b\cdots d}$ and $S^{p\cdots q}{}_{r\cdots s}$. We shall always make this choice. The consequence is that the result of first taking the outer product and then applying tilde is the same as that of first applying tilde to both factors and then taking the outer product. We shall not (because we do not have to) make a separate decision about the assignment of dimension every time we take an outer product.

Again, let $T^{a\cdots c}{}_{b\cdots d}$ be a tensor field on M with some given dimension. What dimension should we assign to $T^{a\cdots m\cdots c}{}_{b\cdots m\cdots d}$ (one contraction)? A natural choice is to assign to $T^{a\cdots m\cdots c}{}_{b\cdots m\cdots d}$ the same dimension as was assigned to $T^{a\cdots c}{}_{b\cdots d}$. We shall always make this choice. The consequence is that the result of first contracting and then applying tilde is the same as that of first applying tilde and then contracting (for contracting annihilates one covariant and one contravariant index). We shall not (because we do not have to) make a separate assignment of dimension every time we contract over indices.

Once again, let $T^{a\cdots c}{}_{b\cdots d}$ be a tensor field on M with some given dimensions. We shall always assign to $T^{a\cdots\ \ \cdots c}{}_{p\ \ b\cdots d} = g_{pq} T^{a\cdots q\cdots c}{}_{b\cdots d}$ the same dimension as was assigned to $T^{a\cdots c}{}_{b\cdots d}$. (Similarly, we assign to $T^{a\cdots c}{}_{b\cdots\ \ p\ \ \cdots d}$ the same dimension as was assign to $T^{a\cdots c}{}_{b\cdots d}$.) Thus, the result of lowering an index of a tensor field (using g_{ab}) and then applying tilde is the same as that of first applying tilde and then lowering the index (using $\tilde{g}_{ab}$). Proof: Lowering an index gives one more covariant index. But $\tilde{g}_{ab} = \Omega^2 g_{ab}$.) Of course, we set $\tilde{g}^{ab} = \Omega^{-2} g^{ab}$, so $\tilde{g}^{ab}\tilde{g}_{bc} = \delta^a{}_c\ (= \tilde{\delta}^a{}_c)$). Thus, raising an index (with g^{ab}) and then applying tilde gives the same result as first applying tilde and then raising the index (with $\tilde{g}^{ab}$).

Finally, we adopt the convention that tensor fields with a tilde have their indices raised and lowered with $\tilde{g}^{ab}$ and $\tilde{g}_{ab}$, while tensor fields without a tilde have their indices raised and lowered (as always) with g^{ab} and g_{ab}. Now everything is

consistent. One must make a choice, for each new tensor field in the space-time, of what its dimension will be. (This choice is based on convenience.) Dimensions add under outer product, and are unchanged under contraction and raising and lowering of indices. The metric itself has dimension zero (which, of course, it must have in order that the dimension remain invariant under raising and lowering of indices). (Note that $\tilde{g}^{ab}$ also has dimension zero.)

It should be emphasized that this scheme is natural and simple. The only freedom we had was the "2" in $\tilde{g}_{ab} = \Omega^2 g_{ab}$, and this choice was based on making formulae as simple as possible.

What dimension should we assign to the alternating tensor, ϵ_{abcd}? Clearly, it should have dimension zero (i.e., $\tilde{\epsilon}_{abcd} = \Omega^4 \epsilon_{abcd}$, for then we shall have $\tilde{\epsilon}_{abcd}\tilde{\epsilon}^{abcd} = -24$. That is, $\tilde{\epsilon}_{abcd}$ will be the alternating tensor for the metric $\tilde{g}_{ab}$.

We are now half done. We have treated the algebraic part of conformal transformations. What remains is the differential part. This requires, not more conventions, but rather a few calculations.

Let ∇_a and $\widetilde{\nabla}_a$ be the derivative operators with respect to g_{ab} and $\tilde{g}_{ab}$, respectively (so $\nabla_a g_{bc} = \widetilde{\nabla}_a \tilde{g}_{bc} = 0$). Then, since these are both derivative operators, there exists a tensor field $C^m{}_{ab} = C^m{}_{(ab)}$ on M such that

$$\begin{aligned}\widetilde{\nabla}_p T^{a\ldots c}{}_{b\ldots d} &= \nabla_p T^{a\ldots c}{}_{b\ldots d} + C^a{}_{pm} T^{m\ldots c}{}_{b\ldots d} + \cdots + \\ &+ C^c{}_{pm} T^{a\ldots m}{}_{b\ldots d} - C^m{}_{pb} T^{a\ldots c}{}_{m\ldots d} - \cdots - C^m{}_{pd} T^{a\ldots c}{}_{b\ldots m}\end{aligned} \tag{144}$$

for every tensor field $T^{a\ldots c}{}_{b\ldots d}$ on M. Thus, this $C^m{}_{ab}$ gives us $\widetilde{\nabla}_a$ in terms of ∇_a. We wish to obtain an explicit expression for $C^m{}_{ab}$. We proceed as follows:

$$\begin{aligned}0 = \widetilde{\nabla}_a \tilde{g}_{bc} &= \nabla_a \tilde{g}_{bc} - C^m{}_{ab}\tilde{g}_{mc} - C^m{}_{ac}\tilde{g}_{bm} \\ &= \nabla_a(\Omega^2 g_{bc}) - 2\,\Omega^2 C^m{}_{a(b} g_{c)m} \\ &= 2\,\Omega g_{bc}\nabla_a\Omega - 2\,\Omega^2 C^m{}_{a(b} g_{c)m}\end{aligned} \tag{145}$$

where, in the second step, we have used (144). But this equation is easy to solve:

$$C^m{}_{ab} = \frac{1}{2}\Omega^{-1} g^{mn}(g_{na}\nabla_b\Omega + g_{nb}\nabla_a\Omega - g_{ab}\nabla_n\Omega) \tag{146}$$

Thus, with the aid of (144) and (146), we can convert from $\widetilde{\nabla}$'s to ∇'s at will, obtaining, by such a conversion, additional terms involving derivatives of Ω.

We give two examples of these remarks. Let T^{ab} be a tensor field of dimension sec^{-24}. Then

$$\begin{aligned}&\widetilde{\nabla}_c \widetilde{T}^{ab} = \nabla_c \widetilde{T}^{ab} + C^a{}_{cm}\widetilde{T}^{mb} + C^b{}_{cm}\widetilde{T}^{am} \\ &= \nabla_c(\Omega^{-6}T^{ab}) + \Omega^{-6}T^{mb}\tfrac{1}{2}\Omega^{-1}(\delta^a{}_c\nabla_m\Omega + \delta^a{}_m\nabla_c\Omega - g_{mc}\nabla^a\Omega) \\ &+ \Omega^{-6}T^{am}\tfrac{1}{2}\Omega^{-1}(\delta^b{}_c\nabla_m\Omega + \delta^b{}_m\nabla_c\Omega - g_{mc}\nabla^b\Omega \\ &= \Omega^{-6}\nabla_c T^{ab} - 5\Omega^{-7}T^{ab}\nabla_c\Omega + \tfrac{1}{2}\Omega^{-7}\delta^a{}_c T^{mb}\nabla_m\Omega \\ &+ \tfrac{1}{2}\Omega^{-7}\delta^b{}_c T^{am}\nabla_m\Omega - \tfrac{1}{2}\Omega^{-7}T_c{}^b\nabla^a\Omega - \tfrac{1}{2}T^a{}_c\nabla^b\Omega\end{aligned}$$

where, in the first step, we have used (144), and, in the second, we have used (146).

As a second example, we evaluate the Riemann tensor, $\widetilde{R}_{abc}{}^d$, of the metric $\tilde{g}_{ab}$. As always, one obtains the Riemann tensor by commuting derivatives. Fix a vector field k_c on M. Then

$$\begin{aligned}
&\tfrac{1}{2}\widetilde{R}_{abc}{}^{d}\,k_d = \widetilde{\nabla}_{[a}\widetilde{\nabla}_{b]}k_c = \nabla_{[a}\widetilde{\nabla}_{b]}k_c - C^{m}{}_{[ab]}\widetilde{\nabla}_m k_c - C^{m}{}_{c[a}\widetilde{\nabla}_{b]}k_m \\
&= \nabla_{[a}\widetilde{\nabla}_{b]}k_c - C^{m}{}_{c[a}\widetilde{\nabla}_{b]}k_m = \nabla_{[a}(\nabla_{b]}k_c - C^{m}{}_{b]c}k_m) \\
&- C^{m}{}_{c[a}(\nabla_{b]}k_m - C^{n}{}_{b]m}k_n) = \nabla_{[a}\nabla_{b]}k_c - k_m\nabla_{[a}C^{m}{}_{b]c} \\
&- C^{m}{}_{c[b}\nabla_{a]}k_m - C^{m}{}_{c[a}\nabla_{b]}k_m + C^{m}{}_{c[a}C^{n}{}_{b]m}k_n \\
&= \tfrac{1}{2}R_{abc}{}^{d}\,k_d - k_d\nabla_{[a}C^{d}{}_{b]c} + C^{m}{}_{c[a}C^{d}{}_{b]m}k_d
\end{aligned}$$

where, in the second step, we have used (144), in the third step, we have used $C^{m}{}_{[ab]} = 0$, in the fourth step, we have again used (144), in the fifth step, we have expanded, and, in the sixth step, we have used the definition of the Riemann tensor. Since k_d is arbitrary, we have

$$\widetilde{R}_{abc}{}^{d} = R_{abc}{}^{d} - 2\nabla_{[a}C^{d}{}_{b]c} + C^{m}{}_{c[a}C^{d}{}_{b]m}$$

One could now substitute (146) and obtain a detailed expression for $\widetilde{R}_{abc}{}^{d}$ in terms of $R_{abc}{}^{d}$ and derivatives of the conformal factor.

It should be clear from the remarks above that, once a dimension has been assigned to each tensor field in sight, every tensor equation can be converted to a tilde-equation, and, conversely, tensor equations with tildes can be converted to equation without tildes.

By using conformal transformation, equations can sometimes be simplified or made more transparent geometrically. A conformal transformation is essentially the only simple thing one can do to a metric. Finally, conformal transformation can be used to bring out structure not apparent from the physical metric. We shall see an example of the latter application in the following sections.

36. Asymptotic Structure: Introduction

It is often the case in physics that one seeks a description of a system in which certain details of its structure are suppressed, and certain overall properties use in the description. (Example: The description of a gas by its pressure, density, and temperature.) Such a global description is possible, in special relativity, essentially because of the action of the Poincaré group as the group of isometries on Minkowski space. In particular, as we have seen, the presence of Killing vectors (the generators of Poincaré group – the "infinitesimal isometries") leads to the notions of the energy, momentum, angular momentum of a closed system in special relativity. Stated in somewhat more vague terms, the symmetry of Minkowski space allows a "comparison of tensors at different space-time points". Hence, one can "add effects from various regions of space-time to obtain a description of our system as a whole".

Unfortunately, in general relativity there are, in the "generic case", no Killing vectors at all – no symmetries. Thus, the special relativistic possibilities for an overall description of a closed system are lost. Can anything at all be done in this direction in curved space? Does there exist an – even partial – global description of a closed system in general relativity?

Such a description is indeed possible if we restrict consideration to space-time which are asymptotically flat. Asymptotic flatness means, (very) roughly speaking, that the metric of space-time approaches a Minkowski metric in the limit as one moves away from the source (i.e., at "infinity") . (In fact, much of the work in this game is directed toward making the previous sentence precise.) For an asymptotically flat space-time, the symmetries (and hence, the structure) of special relativity become valid in the limit at infinity. The global characterization of a system results from examining the approach of the metric to a Minkowski metric. It is through the details of this "approach" that one describes the system. (The situation has some features in common with electrodynamics. The rate of approach of the electric field to zero at infinity is a measure of the total charge of the system.)

One consequence of this approach to describing a system is that the quantities one obtains (e.g., energy) will not be related, in any direct or simple way, to, say, the stress-energy of the system itself. One cannot carry the information at infinity, in any natural way, across the curvature of space-time to compare it with local information about the system being described.

One wants to define what he means by "asymptotically flat". Incorporated into such a definition will be a statement of how quickly the metric must "approach a Minkowski metric". On what rate of approach should one insist? This turns out to be a rather delicate question. We here comment on the competing factors. If the metric becomes flat too slowly, then the asymptotic behavior is not "sufficiently close to that of Minkowski space" that one recovers any global description at all. If, on the other hand, the approach to flatness is too fast, then the asymptotic metric is "so much like that of Minkowski space" that one obtains no information from the behavior of the metric. To put matters another way, a nonzero "total energy" for a system leads to a certain rate of approach of the metric to a Minkowski metric. If, in one's definition of asymptotic flatness, one requires an approach to flatness faster than this rate, then one loses the possibility of even describing systems of nonzero mass (i.e., of describing all interesting systems). Definition of

asymptotic flatness represent a compromise between these two effects.

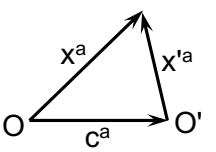
The symmetry group of flat (Minkowski) space is the Poincaré group. An analogous "asymptotic symmetry group" exists for an asymptotically flat space-time. It represents, roughly speaking, the "asymptotic motions under which the asymptotic structure is invariant". The asymptotic symmetry group for general relativity has certain features in common with the Poincaré group, and certain features very different from the Poincaré group. There follows an intuitive discussion of why this group turns out to have the structure it does.

Let us consider again the action of Poincaré group on Minkowski space. The action of the generators of this group is represented by the Killing vectors in Minkowski space. We have seen that the general Killing vector is given by

$$\xi^a = \underline{F}^a{}_b\, x^b + \underline{\xi}^a \tag{147}$$

where $\underline{\xi}^a$ is a constant vector field, $\underline{F}_{ab}$ is a constant antisymmetric tensor field, and x^a is the position vector relative to some origin O of space-time. What happens to (147) when we change our (arbitrary) choice of origin? Let O' be a second origin, and denote by c^a the position vector of O' relative to O. Then the position vectors of a point relative to O and O', x^a and x'^a respectively, are related by $x'^a = x^a - c^a$. Substituting, (147) takes the form $\xi^a = \underline{F}^a{}_b\, x'^b + (\underline{\xi}^a + \xi^a + \underline{F}^a{}_b\, c^b)$.

Denote by P the Poincaré group, and let T be the subgroup consisting of translations (things generated by constant Killing vectors). Then T is normal subgroup of P, and the quotient group, $P/T = L$, the Lorentz group. However, P cannot be written as a product of T and L. This remark is a consequence of the discussion above. A Killing vector defines an origin-independent $\underline{F}_{ab}$, but no origin-independent $\underline{\xi}^a$.



We now consider the asymptotic ("large" x^a) behavior of the Killing vector (147). Clearly, the first term on the right in (147) (the linear term) dominates the second (constant) term. What might we expect an "asymptotically Killing vector" to look like? Firstly, we might demand that it be "asymptotically linear in position", by analogy with (147). One might next be tempted to demand that "after the part asymptotically linear in position is subtracted out, what remains is asymptotically constant". But we cannot demand this! The problem is that "linear in position" is a flat-space notion – one good only in the limit in an asymptotically flat-space time. There arises an ambiguity in separating out a unique "constant part" from an "asymptotic Killing vector" because of the ambiguity in saying what "asymptotically linear in position" means. Of course, as the space-time gets flatter and flatter (as one moves away from the source), this difficulty (which is associated with curvature) diminishes in importance. But the domination of the constant part of (147) by the linear part increases as one moves to infinity. The consequence is that the "constant part" of an "asymptotically Killing vector" simply cannot be given any meaning. We are forced to admit as an "asymptotic Killing vector" any vector field which is asymptotically linear in position. In particular, we may add to such a vector field another which is "bounded", to obtain another vector field asymptotically linear in position. That is to say, the translation subgroup gets enlarged – from constant vector fields in the flat case to "bounded vector field" in the asymptotically flat case. Of course, the Lorentz part of the Poincaré group is unaffected.

Thus, the asymptotic symmetry group of general relativity is an infinite-dimensional group. There is an infinite-dimensional subgroup – called the super translations – which generalizes the finite-dimensional subgroup of translations in the Minkowski case, the quotient group is still the Lorentz group.

The remarks above are intended merely to give an intuitive feel for the structure of the asymptotic symmetry group in general relativity. In practice, the group simply pops up: it is neither introduced nor discussed in these terms.

Asymptotic structure deals with the limit at infinity. One might have thought, therefore, that the description will involve limiting procedures. Limits, of course, are inconvenient. A method has been devised for overcoming them. The reason why limits arise, of course, is that "infinity is far away". So, one "brings infinity in to a finite place, and represents it by additional points attached to space-time". How can one bring in something far away? By a conformal transformation. Thus, one introduces an unphysical metric $\tilde{g}_{ab} = \Omega^2 g_{ab}$. In order that we "shrink distances" to "bring in infinity," we choose Ω to approach zero asymptotically. With sufficient care, one can arrange matters so that the space-time manifold M, with metric admits a boundary surface of points "at infinity". Then asymptotic structure is described in terms of local structure at these points at infinity. One merely applies a conformal transformation to the relevant equations, and sees what structure emerges at the points at infinity. In fact, one goes one step further: asymptotic flatness is defined by the existence of an appropriate conformal transformation which allows a certain collection of points at infinity.

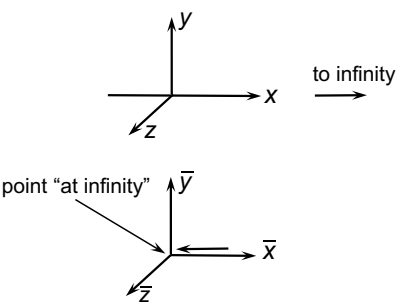
This conformal-transformation-points-at-infinity approach is convenient because it allows one to avoid complicated (and usually imprecise) limiting procedures. Put another way, one obtains the possibility of bringing local differential geometry ("local" at the points of infinity) to bear on asymptotic problems. Differential geometry replaces limits. It seems always to turn out in practice that, when something can be formulated in terms of differential geometry, it becomes simple and transparent.

In fact, there are two regimes in which one discusses asymptotic structure in general relativity: at spatial infinity and null infinity. (They refer to the mode of approach to infinity.) We shall discuss some of the special features of each regime in the following sections. One important and unsolved problem in general relativity is to relate the description of a system at spatial infinity to the description of the same system at null infinity. Essentially nothing is known. It is not even clear that asymptotic flatness in one regime implies asymptotic flatness in the other. The remarks above have been collected together because they apply to both regimes.

37. Asymptotic Structure at Spatial Infinity

We now consider the asymptotic structure of the gravitational field in the limit as one moves off to infinity "spatially". From the discussion of signal propagation in Sect. 28, one would expect that no effects of anything which happens in a compact region of space-time could have any influence on spatial infinity. Thus, the structure at spatial infinity at one "instant of time" completely determines that structure thereafter.

Consider initial data for our space-time. That is, we have a three-dimensional manifold S with positive-definite metric h_{ab} and extrinsic curvature Π^{ab}. The description of the asymptotic structure of our space-time at spatial infinity begins with the description of the asymptotic structure of this initial-data set. Roughly speaking, we wish to require that (S, h_{ab}) "approach Euclidean space" asymptotically, while Π^{ab} approaches zero. To see how this statement can be formulated precisely, we must discuss some properties of Euclidean space.



Consider Euclidean space, with the usual coordinates, x, y, z, in terms of which the metric is $(\mathrm{d}x)^2 + (\mathrm{d}y)^3 + (\mathrm{d}z)^2$. Set $r^2 = x^2 + y^2 + z^2$), and introduce new coordinates

$$\bar{x} = x/r^2, \quad \bar{y} = y/r^2, \quad \bar{z} = z/r^2, \tag{148}$$

Note that the "point" $\bar{x} = \bar{y} = \bar{z} = 0$ corresponds to "infinity" ($x \to \infty,\ y \to \infty,\ z \to \infty$). Thus, approaching infinity along the x-axis (i.e., $x \to \infty,\ y = z = 0$) is, expressed in terms of barred coordinates, approaching $\bar{x} = \bar{y} = \bar{z} = 0$ along the $\bar{x}$-axis (i.e., $\bar{x} \to \infty,\ \bar{y} = \bar{z} = 0$). Thus, (148) suggests that we add a single point "at infinity" to Euclidean space, with $\bar{x}, \bar{y}, \bar{z}$ a coordinate system valid in a neighborhood of this point.

The next step is to rewrite the metric in terms of this barred coordinate system. First note that $\mathrm{d}\bar{x} = (y^2+z^2-x^2)\,\mathrm{d}x/r^4 - 2xy\,\mathrm{d}y/r^4 - 2xz\,\mathrm{d}z/r^4$, etc. for $\mathrm{d}\bar{y},\ \mathrm{d}\bar{z}$. Hence,

$$r^4(\,\mathrm{d}\bar{x}^2 + \,\mathrm{d}\bar{y}^2 + \,\mathrm{d}\bar{z}^2) = (\,\mathrm{d}x^2 + \,\mathrm{d}y^2 + \,\mathrm{d}z^2) \tag{149}$$

Thus, if we wish to regard $\bar{x}$, $\bar{y}$, $\bar{z}$ as "good" coordinates near infinity, then the original metric of our Euclidean space becomes badly behaved near "infinity", $\bar{x} = \bar{y} = \bar{z} = 0$. This, of course, is to be expected. After all, infinity is a long way away (according to the original Euclidean metric). If we insist that there be a point at infinity, then, in order that this point be infinitely far away from other points of our space, it is necessary that the Euclidean metric grow without bound near this point at infinity.

Denote by h_{ab} the original Euclidean metric. Then, from (149), the metric $\tilde{h}_{ab} = \Omega^2 h_{ab}$ is well-behaved near infinity (in the coordinates valid in that region: $\bar{x}$, $\bar{y}$, $\bar{z}$), provided we choose $\Omega = 1/r^2$. This, too, is what we might have expected. In order to obtain a metric which is well-behaved near infinity, we must perform

a conformal transformation using a conformal factor Ω which approaches zero at infinity. In this way, we "shrink distances", bringing infinity in to a finite distance from other points of our space.

How shall we express, in a coordinate-independent way, the asymptotic behavior of the conformal factor, $\Omega = 1/r^2$? Set $\bar{r}^2 = \bar{x}^2 + \bar{y}^2 + \bar{z}^2$. Then, from (148), $\Omega = \bar{r}^2$. Next, note that $\bar{r}$ represents, geometrically, the distance from infinity, as measured by the metric $\tilde{h}_{ab}$. Hence, at the point at infinity, $\Omega = 0$, $\widetilde{D}_a\Omega = 0$, and $\widetilde{D}_a\widetilde{D}_b\Omega = 2\tilde{h}_{ab}$, where $\widetilde{D}_a$ is the derivative (with respect to $\tilde{h}_{ab}$). It is clear that these conditions describe the asymptotic behavior of the conformal factor.

The discussion above can be regarded as motivation for the simple and precise definition which follows. We have deliberately been sloppy in our discussion of Euclidean space, for the asymptotic structure of Euclidean space, suggested above, is expressed very neatly by our definition. The idea is to define asymptotic flatness of a space by the existence of a "conformal completion" which has the essential properties of the conformal completion described above for Euclidean space.

Let S be a three-dimensional manifold with positive-definite metric h_{ab}. This (S, h_{ab}), will be said to be *asymptotically flat* if there exists a manifold $\widetilde{S} = S \cup \Lambda$, where Λ is an abstract point (i.e., S is a sub manifold of $\widetilde{S}$, with $\widetilde{S} - S$ a single point), with metric $\tilde{h}_{ab}$ (on $\tilde{S}$) such that

1. On S, $\tilde{h}_{ab} = \Omega^2 h_{ab}$, where Ω is a scalar field on S.
2. At the point Λ, $\Omega = 0$, $\widetilde{D}_a\Omega = 0$, and $\widetilde{D}_a\widetilde{D}_b\Omega = 2\tilde{h}_{ab}$.

(Note: In our discussion of Euclidean space, the conformal factor, $1/r^2$, became singular at the origin. Clearly, the same discussion could have been carried out – less conveniently – using a conformal factor which is well-behaved at the origin.)

We now claim that Euclidean 3-space is asymptotically flat in the sense above. It should be clear that asymptotic flatness is a reasonable replacement for the intuitive statement that the metric "becomes Euclidean sufficiently quickly at infinity".

We now return to our initial data set, (S, h_{ab}, Π^{ab}). We want to define asymptotic flatness for such a thing. We first require that (S, h_{ab}), be asymptotically flat in the sense above. Intuitively, this represents a condition on the "geometry". What remains is to impose condition on the "fields", including, in particular, "Π^{ab}". It turns out that the appropriate conditions are that the tensor fields $\widetilde{\Pi}^{ab}$, $\Omega^{1/2}\widetilde{R}_{ab}$, and $\Omega^{-1/2}(\widetilde{D}_a\widetilde{D}_b\Omega - 2\tilde{h}_{ab})$ have limits at the point Λ, where we have assigned to Π^{ab} dimensions sec^{-1} (i.e., $\widetilde{\Pi}^{ab} = \Omega^{-3}\Pi^{ab}$). Thus, in particular, we require that Π^{ab} approach zero at infinity, at a certain rate. The values of these three tensors at the point Λ at infinity represent the asymptotic gravitational field. The asymptotic quantities characteristic of our system as a whole are defined in terms of these three tensors.

This, then, is the framework of the discussion of asymptotic structure at spatial infinity. One chooses an initial-data surface which is asymptotically flat, introduces a point Λ at infinity, and obtain certain tensors at that point which describe the asymptotic behavior of the gravitational field.

One is interested, of course, in the asymptotic structure of space-time, not in the asymptotic structure of "space". In other words, one would want the description above to be independent of the choice of initial surface – or, at least, to have a simple dependence on the choice of surface. We next discuss how one investigates dependence on the choice of surface.

The equations which govern the evolution of initial data are (136) and (137), where φ, the evolution function, is at our disposal. It describes the distance one moves (normally) off the initial surface S at each point of S. Suppose our initial (S, h_{ab}, Π^{ab}), were asymptotically flat. It is clear, at least intuitively, that for

certain choices of φ, the next surface will not be asymptotically flat (e.g., choose φ to become infinite, quickly, near infinity). Thus, in order that asymptotic flatness (of $(S,\, h_{ab},\, \Pi^{ab})$) be preserved during the evolution, we must be selective in our choice of φ. What conditions must we impose on φ in order that asymptotic flatness be preserved? To find out, we return to the case of Minkowski space.

Suppose that our initial S were a three-dimensional plane in Minkowski space, i.e., suppose h_{ab} were flat and Π^{ab} vanished. For what choices of φ are the succeeding surfaces also planes? The answer is immediate from (136) and (137): we must have $D^a D^b \varphi = 0$. In other words, φ must have the form

$$\varphi = s_a x^a + s \tag{150}$$

where s_a is a constant vector field, s is a constant scalar field, and x^a is the position vector relative to some origin (in our Euclidean space). Geometrically, $\varphi = s =$ constant, corresponds to an infinitesimal time-translation of our surface, while $\varphi = s_a x^a$ corresponds to an infinitesimal boost. Thus, in the general curved case, we must demand that, asymptotically, φ behave like (150). We must characterize this asymptotic behavior locally, in term of behavior at Λ. Let φ have dimensions sec, so $\widetilde{\varphi} = \Omega\varphi$. Then (150) gives $\widetilde{\varphi} = r^{-2} s_a x^a + r^{-2} s = s_a \tilde{x}^a + \tilde{r}^2 s$. Hence, at the point Λ,

$$\widetilde{\varphi} = 0 \quad \widetilde{D}_a \widetilde{\varphi} = s_a \quad \widetilde{D}_a \widetilde{D}_b \widetilde{\varphi} = 2\, s\, \tilde{h}_{ab} \tag{151}$$

In words, $\tilde{\varphi}$ vanishes at Λ, while the derivative of $\widetilde{\varphi}$ at Λ gives the amount of boost represented by φ, and the second derivative of $\widetilde{\varphi}$ at Λ is a multiple of the metric, the multiplicative factor giving the amount of time-translation represented by φ.

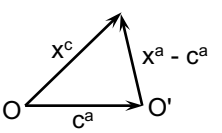
Thus, to preserve asymptotic flatness in the curved case, one would be tempted to insist that φ be such that, at Λ, the conditions (151) hold. It turns out that, whereas the first two conditions (151) can indeed be imposed in the presence of curvature, the third cannot. This remark is suggested by the following equation

$$\widetilde{D}_{[a} \widetilde{D}_{b]} \widetilde{D}_c \widetilde{\varphi} = \tfrac{1}{2} \widetilde{\mathscr{R}}_{abc}{}^{d}\, \widetilde{D}_d\, \widetilde{\varphi} = \tfrac{1}{2} \widetilde{\mathscr{R}}_{abc}{}^{d}\, s_d \tag{152}$$

where, in the second step, we have used the second equation (151). It is perhaps not surprising, from (152), that, when $\widetilde{\mathscr{R}}_{abcd} = 0$ and $s_a = 0$, we are not free to choose $\widetilde{D}_a \widetilde{D}_b \widetilde{\varphi}$, at Λ, to be a multiple of $\tilde{h}_{ab}$. In words, in the presence of curvature and a boost, time-translations can no longer be cleanly and unambiguously distinguished. What one does, in light of (152) is to enlarge the time-translations to encompass possibilities more general than $\widetilde{D}_a \widetilde{D}_b \widetilde{\varphi} = 2\, s\, \tilde{h}_{ab}$. Here is the "enlargement of the translation subgroup" which appears in the asymptotic symmetry group – the phenomenon discussed in Sect. 36.

We have now discussed boosts (which are represented by vectors at Λ, just as in the flat case, and which, therefore, play the same role as in flat space), time-translations (which, because of (152), must be enlarged to an infinite-dimensional group). Spatial rotations, of course, are merely represented by rotations within the tangent space of Λ (The reason: Perform a rotation on x, y, z in Euclidean space, and see what it does to $\bar{x}, \bar{y}, \bar{z}$ and hence to the tangent space at Λ.) Where do the spatial translations show up in the formalism?

To answer this question, we return to the Euclidean case. Recall that we choose for our conformal factor $\Omega = 1/r^2$, where r is the distance from some origin. A translation (spatial) can be represented by a change in this choice of origin and, hence, by a change in choice of conformal factor. Let O' be a second



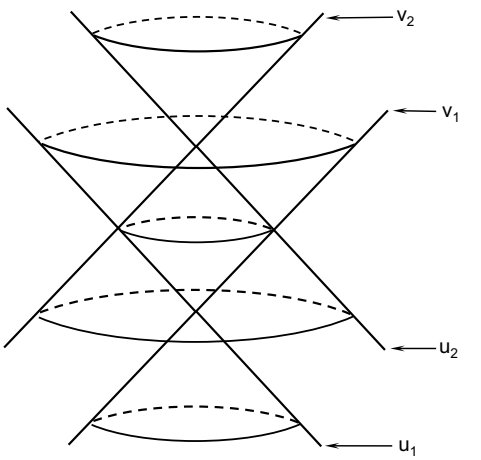
origin, with position vector c^a with respect to O. Then, writing r' for the distance from O' we have $(r')^2 = (x^a - c^a)(x_a - c_a) = r^2 - 2c_a x^a + c^a c_a$. Hence, the conformal factor which results from choosing O' as our origin is $\Omega' = (r^2 - 2c_a x^a + c^a c_a)^{-1}$. Set $\Omega = \Omega'\omega$, so $\omega = 1 - 2c_a x^a r^{-2} = 1 - 2c_a \bar{x}^a + (c_a c^a)\bar{r}^2$ That is to say, the transition from conformal factor Ω to conformal factor Ω' is is accomplished by a conformal transformation with conformal factor $\omega = 1 - 2c_a \bar{x}^a + c_a c^a \bar{r}^2$. This ω is one at Λ, while its derivative at Λ represents the spatial translation (i.e., c^a). Similarly, the behavior of fields at Λ in the general (curved) case under further conformal transformations represents the effects of spatial translations.

The remarks above are intended as a survey of what is, in detail, a rather technical subject. We summarize. Asymptotic structure at spatial infinity is described in terms of an initial-data set for the space-time. We attach to this three-dimensional manifold a single point Λ at infinity, and extend the metric to this point by means of a conformal transformation. Asymptotic flatness of our initial data set is defined by the existence of such a conformal completion, such that certain tensor fields assume finite values at Λ. These tensors at Λ describe the asymptotic gravitational field. Asymptotic boosts and time-translations show up in the asymptotic behavior of the evolution function φ. Asymptotic rotations show up as rotations of the tangent space of Λ. Asymptotic spatial translation show up as conformal transformations. All these motions, taken together, constitute the asymptotic symmetry group at spatial infinity.

38. Asymptotic Structure at Null Infinity

The second regime for the study of asymptotic structure is in the limit at null infinity. We briefly summarize the set-up.

Consider Minkowski space in spherical coordinates: $-\,\mathrm{d}t^2+\,\mathrm{d}r^2+r^2(\,\mathrm{d}\theta^2+\sin^2\theta\,\mathrm{d}\varphi^2)$. Replace the coordinates r, t by new coordinates $u = t + r$ and $v = t - r$, so the metric becomes $-\,\mathrm{d}u\,\mathrm{d}v + (1/4)(u-v)^2(\,\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2)$. Geometrically, the surfaces of constant u are past null cones, centered on the t-axis, while the surfaces v = const, are future-directed null cones. The two limits with which we shall be concerned are future null infinity ($u \to \infty$, with v, θ, φ finite) and past null infinity ($v \to -\infty$, with u, θ, and φ finite)

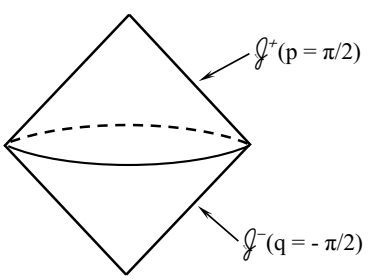


The idea is to do here for Minkowski space what we did, in Sect. 37, for Euclidean space. We cleverly select a system of coordinates in terms of which there is a natural conformal transformation on our metric and a natural "boundary at infinity". We then describe the situation in coordinate-independent terms. Finally, we use the existence of appropriate conformal completion as the definition of asymptotic flatness.

We must introduce still another coordinate transformation in Minkowski space such that our asymptotic limits ($u \to \infty$, with v, θ, φ finite; $v \to -\infty$, with u, θ, φ finite) appear at finite coordinate values. Set $u = \tan p$, $v = \tan q$. Then $u \to \infty$ corresponds to $p = \pi/2$, and $v \to$ to $q = -\pi/2$. In terms of these coordinates, our metric takes the form

$$\cos^{-2} p\, \cos^{-2} q\, [-\,\mathrm{d}p\,\mathrm{d}q + \tfrac{1}{4}\sin^2(p-q)(\,\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2)]$$

Geometrically, we have a null, three-dimensional surface $\mathscr{J}^+(p = \pi/2)$, and a second null, three-dimensional surface $\mathscr{J}^-(q = -\pi/2)$ attached to our space-time. These are the surfaces at "null infinity." Asymptotic structure at null infinity is to be describes as local structure on these surfaces. Our metric, (113) is not, of course, well-behaved on $\mathscr{J}^+$ and $\mathscr{J}^-$. However, $\Omega^2 g_{ab}$ is well-behaved, where $\Omega = \cos p \cos q$. Note that the conformal factor Ω vanishes at infinity (as it must, in order to bring in infinity to a finite surface). Finally, note that, on $\mathscr{J}^\pm$, $\widetilde{\nabla}_a\Omega$, is null and nonzero.



These observations about Minkowski space having been made, we are prepared to define asymptotic flatness of a space-time at null infinity. Let (M, g_{ab}) be a space-time. This space-time is said to be asymptotically flat (at null infinity) if there exists a manifold $\widetilde{M} = M \cup \mathscr{J}^+ \cup \mathscr{J}^-$, where $\mathscr{J}^+$ and $\mathscr{J}^-$ are boundary surfaces (more formally, we should introduce a precise notion here: that of

manifold-with-boundary), which metric $\widetilde{g}_{ab}$ such that

1. On M, $\widetilde{g}_{ab} = \Omega^2 \, g_{ab}$, where Ω is some scalar field on $\widetilde{M}$.
2. $\mathscr{J}^{\pm}$ are null surfaces.
3. On $\mathscr{J}^{\pm}$, $\Omega = 0$, and $\widetilde{\nabla}_a \Omega$ is a nonzero null vector.

We now assert, from our earlier remarks, that Minkowski space is asymptotically flat in this sense.

How does the description of asymptotic structure at null infinity compare with that at spatial infinity? We remarked earlier that nothing ever happens at spatial infinity. Quite the reverse is the case at null infinity. From our remarks on signal propagation, it is clear that radiation escapes to null infinity: it is registered on $\mathscr{J}^{+}$. Thus, null infinity (the "radiation zone") is the natural asymptotic limit for the discussion of radiation. On the other hand, conservation laws are more easily discussed at spatial infinity.

There are numerous unsolved problems in general relativity concerning asymptotic structure. Perhaps the most interesting are those concerning the relation between the two asymptotic descriptions. Essentially all questions of the type "If such-and-such is true at special infinity, does it follow that such-and-such is true at null infinity?" are unanswered. It is not even known whether not asymptotic flatness in one of the two senses implies asymptotic flatness in the other!

39. Linearization

To what does Einstein's equation reduce when the metric of space-time is "nearly flat". More precisely, suppose we set $g_{ab} = \eta_{ab} + \gamma_{ab}$, where η_{ab} is a Minkowski (flat) metric, and γ_{ab} is "small". We could then imagine expanding Einstein's equation to first order in the perturbation, γ_{ab}. The resulting equations would provide an approximate description of "weak gravitational fields" (i.e., fields sufficiently weak that they do not require that the geometry depart significantly from flatness). These equations provide a simple picture of the behavior of gravitational fields in this limit. They are called the linearized Einstein equation. We obtain the linearized equation in the present section.

It is convenient to adopt an approach in which one does not have to keep track of the orders of various terms in the perturbation expansion. Let $g_{ab}(\lambda)$ be a one-parameter family of metrics on a fixed manifold M, where the parameter λ has range, say $[0, 1)$. Let $g_{ab}(0)$ be a flat metric (i.e., with vanishing Riemann tensor). The plan is to evaluate $\mathrm{d}/\,\mathrm{d}\lambda$ of various quantities at $\lambda = 0$. Thus, by introducing this "expansion parameter λ," and taking derivatives with respect to λ, we automatically keep the appropriate terms in our perturbation expansion.

For each value of λ, denote by $\overset{\lambda}{\nabla}_a$ the derivative operator defined by the metric $g_{ab}(\lambda)$. Fix a ($\lambda-$independent) vector field k_b on M. Then

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\overset{\lambda}{\nabla}_a k_b\right)\Big|_{\lambda=0} = -C^m{}_{ab}\, k_m \tag{153}$$

for some tensor field $C^m{}_{ab} = C^m{}_{(ab)}$. (Proof: Note that the left side of (153) is linear in k_b, and that the anti symmetric part of the left side vanishes.) Our first task is to evaluate this $C^m{}_{ab}$. Setting $g_{ab} = g_{ab}(0)$, $\nabla_a = \overset{0}{\nabla}_a$, we have

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\overset{\lambda}{\nabla}_a g_{bc}(\lambda)\right]\Big|_{\lambda=0} \\ &= -C^m{}_{ab}\, g_{mc} - C^m{}_{ac}\, g_{bm} + \nabla_a\left[\frac{\mathrm{d}}{\mathrm{d}\lambda} g_{bc}(\lambda)\,\Big|_{\lambda=0}\right] \end{aligned} \tag{154}$$

where, in the first step, we have used the fact that $\overset{\lambda}{\nabla}_a\, g_{bc}(\lambda)$ vanishes for all λ, and, in the second step, we have used (153). Set $\gamma_{ab} = (\,\mathrm{d}/\,\mathrm{d}\lambda) g_{ab}(\lambda)|_{\lambda=0}$. Then the solution of (154) is

$$C^m{}_{ab} = \tfrac{1}{2} g^{mn}(\nabla_a \gamma_{bn} + \nabla_b \gamma_{an} - \nabla_n \gamma_{ab}) \tag{155}$$

Think of $C^m{}_{ab}$ as representing the "$\lambda-$rate of change of the derivative operator, evaluated at $\lambda = 0$". This $C^m{}_{ab}$ is given by (155).

For each λ, we have a Riemann tensor, $R_{abc}{}^d(\lambda)$. Of course, $R_{abc}{}^d(0)$ vanishes, for we choose our metric to be flat at $\lambda = 0$. However, the linearized Riemann tensor, $(\,\mathrm{d}/\,\mathrm{d}\lambda) R_{abc}{}^d(\lambda)|_{\lambda=0}$, will not vanish in general. This quantity represents the curvature which results, to first order, from our perturbation of the metric away from a flat metric. To evaluate it, we proceed as follows. Let k_c be a fixed vector field on M. Then, for each value of λ, $\overset{\lambda}{\nabla}_{[a}\overset{\lambda}{\nabla}_{b]} k_c = \frac{1}{2} R_{abc}{}^d(\lambda)\, k_d$. Taking $\mathrm{d}/\,\mathrm{d}\lambda$, and evaluating at $\lambda = 0$,

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\lambda}R_{abc}{}^{d}(\lambda)\,\big|_{\lambda=0}\;k_d = \left[\frac{\mathrm{d}}{\mathrm{d}\lambda}\overset{\lambda}{\nabla}_{[a}\overset{\lambda}{\nabla}_{b]}\,k_c\right]_{\lambda=0}\\
&= -C^{d}{}_{[ab]}\overset{\lambda}{\nabla}_d\,k_c\,\big|_{\lambda=0} - C^{d}{}_{c[a}\overset{\lambda}{\nabla}_{b]}k_d\,\big|_{\lambda=0} + \nabla_{[a}\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\overset{\lambda}{\nabla}_{b]}k_c\,\big|_{\lambda=0}\right)\\
&= -C^{d}{}_{[ab]}\nabla_d k_c - C^{d}{}_{c[a}\nabla_{b]}k_d + \nabla_{[a}(-C^{d}{}_{b]c}k_d)\\
&= -C^{d}{}_{c[a}\nabla_{b]}k_d - \nabla_{[c}C^{d}{}_{b]c}k_d - C^{d}{}_{c[b}\nabla_{a]}k_d\\
&= -k_d\nabla_{[a}C^{d}{}_{b]c}
\end{aligned}$$

where, in the second step, we have used (153), in the third step, we have again used (153), in the fourth step, we have expanded and used $C^{d}{}_{[ab]} = 0$ and, in the fifth step, we have canceled the first and last terms. Since k_d is arbitrary,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}R_{abc}{}^{d}(\lambda)\,\big|_{\lambda=0} = -2\nabla_{[a}C^{d}{}_{b]c} \tag{156}$$

Now substitute (155). (We shall hereafter raise and lower indices with g_{ab}, the "unperturbed metric".)

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}R_{abc}{}^{d}(\lambda)\,\big|_{\lambda=0} = \nabla^{d}\nabla_{[a}\gamma_{b]c} - \nabla_{c}\nabla_{[a}\gamma_{b]}{}^{d} \tag{157}$$

This is the expression, to first order, for the curvature which results from a perturbation of the metric of flat space. Contracting (157) once, (with $\gamma = \gamma^{m}{}_{m}$)

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}R_{ab}(\lambda)\,\big|_{\lambda=0} = \nabla_{(a}\nabla^{m}\gamma_{b)m} - \frac{1}{2}\nabla^{m}\nabla_{m}\gamma_{ab} - \frac{1}{2}\nabla_{a}\nabla_{b}\gamma \tag{158}$$

Now we suppose our family of metrics, $g_{ab}(\lambda)$, represents a family of solutions of Einstein's equation, so we have a stress-energy, $T_{ab}(\lambda)$. Of course, $T_{ab}(\lambda) = 0$, since our initial metric is flat. The quantity $(\,\mathrm{d}/\,\mathrm{d}\lambda)T_{ab}(\lambda)|_{\lambda=0}$ thus represents the "first-order stress-energy of matter". From (158), Einstein's equation becomes

$$\begin{aligned}
8\pi G\frac{\mathrm{d}}{\mathrm{d}\lambda}T_{ab}(\lambda)\,\Big|_{\lambda=0} &= \nabla_{(a}\nabla^{m}\gamma_{b)m} - \frac{1}{2}\nabla^{m}\nabla_{m}\gamma_{ab} - \frac{1}{2}\nabla_{a}\nabla_{b}\gamma\\
&\quad - \frac{1}{2}g_{ab}\nabla^{m}\nabla^{n}\gamma_{mn} + \frac{1}{2}g_{ab}\nabla^{m}\nabla_{m}\gamma
\end{aligned} \tag{159}$$

Eqn, (159) can be simplified somewhat. The steps are very closely analogous to corresponding steps in the discussion of potentials for electromagnetism, so we begin with that case. Recall equations: $\nabla^{[a}F^{bc]} = 0$, $\nabla_b F^{ab} = J^a$. We restrict consideration to flat space. The first equation above implies that

$$F^{ab} = \nabla^{[a}A^{b]} \tag{160}$$

for some vector field A^a (called the *vector potential*) in our flat space. Substituting, the second Maxwell equation gives

$$-\frac{1}{2}\nabla^{m}\nabla_{m}A^{a} + \frac{1}{2}\nabla^{a}(\nabla_{m}A^{m}) = J^{a} \tag{161}$$

Eqn. (160) does not determine this potential A^a uniquely from F^{ab}. In fact, A^a and $A'^{a} = A^{a} + \nabla^{a}\psi$, where ψ is any scalar field, which defines, via (160), the same F^{ab}. This $A'^{a} = A^{a} + \nabla^{a}\psi$ is called a *gauge transformation* on the vector potential. We have

$$\nabla_{a}A'^{a} = \nabla_{a}A^{a} + \nabla_{a}\nabla^{a}\psi \tag{162}$$

Setting the left side of (162) equal to zero, we obtain a wave equation with source for ψ. This equation has a solution. Thus, by means of a gauge transformation, we can obtain a vector potential A^a satisfying (160) and also

$$\nabla_a A^a = 0 \tag{163}$$

Eqn. (163) is called the *Lorentz gauge condition*. Thus, we utilize the freedom available in gauge transformations to obtain a vector potential satisfying the subsidiary condition (163). The advantage is that (161) now simplifies to

$$-\frac{1}{2}\nabla^m \nabla_m A^a = J^a \tag{164}$$

The electromagnetic situation above is completely analogous to the gravitational situation. We regard the vector potential A^a as analogous to the "gravitational potential", γ_{ab}. The electromagnetic field, F^{ab}, is analogous to the left side of (157) – the linearized Riemann tensor. Then (160) is analogous to (157). The electromagnetic charge-current, J^a, is analogous to the linearized stress-energy, $(\,\mathrm{d}/\,\mathrm{d}\lambda)T_{ab}(\lambda)|_{\lambda=0}$. (That is, the source for the electromagnetic field is analogous to the source for the gravitational field.) Thus, (159) is analogous to (161).

What in the gravitational case is analogous to the electromagnetic gauge transformations and Lorentz gauge? First note that, replacing γ_{ab} on the right in (157) by $\gamma'_{ab} = \gamma_{ab} + \nabla_{(a}\xi_{b)}$, where ξ_b is any vector field, the right side of (157) is unchanged. Thus, we regard $\gamma'_{ab} = \gamma_{ab} + \nabla_{(a}\xi_{b)}$ as representing "gravitational gauge transformations", with ξ_b analogous to ψ. These gravitational gauge transformations change the linearized metric (γ_{ab}) without changing the linearized curvature. Next, note that

$$\nabla_b(\gamma'^{ab} - \tfrac{1}{2}\gamma' g^{ab}) = \nabla_b(\gamma^{ab} - \tfrac{1}{2}\gamma g^{ab}) + \tfrac{1}{2}\nabla^m \nabla_m \xi^a \tag{165}$$

Thus, we can always perform a gauge transformation such that

$$\nabla_b(\gamma^{ab} - \tfrac{1}{2}\gamma g^{ab}) = 0 \tag{166}$$

Eqn. (159) then becomes

$$8\pi G \frac{\mathrm{d}}{\mathrm{d}\lambda} T_{ab}(\lambda)\,\Big|_{\lambda=0} = -\frac{1}{2}\nabla^m \nabla_m(\gamma_{ab} - \frac{1}{2}\gamma g_{ab}) \tag{167}$$

We regard (165), (166), and (167) as analogous to (162), (163), and (164), respectively. Note that (167) is essentially a wave equation, with the linearized matter stress-energy acting as a source for the linearized metric.

We summarize with the table below:

The field is	$\frac{\mathrm{d}}{\mathrm{d}\lambda}R_{abcd}\vert_{\lambda=0}$	F_{ab}
We introduce a potential	γ_{ab}	A_a
in terms of which the field is	$\nabla_d\nabla_{[a}\gamma_{b]c} - \nabla_c\nabla_{[a}\gamma_{b]d}$	$\nabla_{[a}A_{b]}$
The potential is determined by the field only up to	$\gamma_{ab} \to \gamma_{ab} + \nabla_{(a}\xi_{b)}$	$A_a \to A_a + \nabla_a\psi$
For arbitrary	ξ_b	ψ
These are the gauge transformations. One can by means of appropriate gauge transformations choose a potential satisfying	$\nabla_b(\gamma^{ab} - \frac{1}{2}\gamma g^{ab}) = 0$	$\nabla_a A^a = 0$
The source for the field is	$\frac{\mathrm{d}}{\mathrm{d}\lambda}T_{ab}(\lambda)\vert_{\lambda=0}$	J^a

The source "generates potential" via	$8\pi G\frac{\mathrm{d}}{\mathrm{d}\lambda}T_{ab}(\lambda)\Big\vert_{\lambda=0} = \nabla_{(a}\nabla^m\gamma_{b)m} - \frac{1}{2}\nabla^m\nabla_m\gamma_{ab} - \frac{1}{2}\nabla_a\nabla_b\gamma - \frac{1}{2}g_{ab}\nabla^m\nabla^n\gamma_{mn} + \frac{1}{2}g_{ab}\nabla^m\nabla_m\gamma$	$J_a = -\frac{1}{2}\nabla^m\nabla_m A_a + \frac{1}{2}\nabla_a(\nabla^m A_m)$
If, however, the gauge condition is satisfied, this simplifies to	$8\pi G\frac{\mathrm{d}}{\mathrm{d}\lambda}T_{ab}(\lambda)\Big\vert_{\lambda=0} = -\frac{1}{2}\nabla^m\nabla_m(\gamma_{ab} - \frac{1}{2}\gamma g_{ab})$	$J_a = -\frac{1}{2}\nabla^m\nabla_m A_a$

The advantage of introducing potentials is in both cases the same. One in this way obtains equations of the form $\nabla^m\nabla_m$ (potentials) = sources. Such equations are easy to solve, e.g., by Green's functions.

40. Quantization of Gravitational Field: Introduction

We have been discussing the general theory of relativity. Perhaps the three fundamental principles of this theory are the following; i) The events of space-time (all events, past, present, and future) are assembled into a four-dimensional manifold. The description of physics is in terms of fields on this manifold. ii) There is a metric tensor field g_{ab}, on this manifold. The metric simultaneously describes the geometry (results of space and time measurements) of space-time and the effect of gravitation. iii) Matter in space-time produces a certain tensor field which causes the metric of space-time to exhibit curvature. One of the central features of the general theory of relativity (a feature we have, perhaps, not stressed strongly enough) is that the theory claims to incorporate within its structure all of physics. Where there's "physics", there's stress-energy, and hence there's curvature of space-time. The full apparatus of general relativity must, at least in principle (though almost never in practice!) be brought into play in the discussion of any physical phenomenon. General relativity claims a universality over other areas of physics. (Note the word "claims". The entire theory could, of course, differ substantially from the way Nature chooses to behave.)

There exists at least one other theory of physics with a similar claim to universality: quantum theory. In my opinion, the fundamental principles of quantum theory are: i) The states of a system are described in terms of a Hilbert space (more specifically, by rays in a Hilbert space), and ii) the attributes (properties of, measurements on, etc.) of the system are described in terms operators on that Hilbert space. Perhaps quantum theory can be viewed as the insistence that every theory of physics be formulated according to the principles above.

That there is a problem here should now be clear. The general theory of relativity is not formulated in the terms demanded by quantum theory. One seeks, therefore, a modification of the theory to obtain consistency with the principles of quantum theory: one seeks to quantize the general theory of relativity. This new theory should, presumably, not go any more strongly than necessary against the fundamental principles of general relativity. The problem, then, is to write down a theory which is both "quantum-theory-looking" and "general-relativity-looking". This problem, to which a great deal of effort and many clever ideas have been directed, remains unsolved. We shall, in the next few sections, discuss a few of the approaches to this problem. It should be emphasized that whether or not a collection of sentences and equations represents a "solution" to this problem is, for the most part, an aesthetic question.

What features might one expect to appear in a quantum theory of gravitation? It is common that concepts in a classical theory which are "sharp" become "fuzzed out" on quantization. For example, for a particle approaching a potential barrier, classically, the particle either reflects or is transmitted, while, in quantum theory, there is merely a distribution in probabilities for various outcomes. The primary candidate for something to be fuzzed out on quantization of general relativity is the point events of space-time. One might expect that these events will lost their significance – i.e., will reappear only in the classical limit of the quantum theory of gravitation. This expectation is suggested, for example, by the following remark. Suppose we build a probe of some sort which makes measurements in a very small region of space-time (or, in the limit, at a single event of space-time). Then our probe must be at least as small as the region over which it

makes measurements. But the uncertainty principle in quantum theory suggests that very small instruments must contain particles of high momentum – hence, high energy. But, if our probe is to have a large stress-energy, then by Einstein's equation, it must be responsible for large curvatures of space-time. In other words, a significant distortion of space-time in the region being measured will result from introducing our probe. Thus, it appears that point events will lose their operational significance under quantization. But it usually happens in physics that, when a concept looses operational significance, that loss is reflected in the mathematical formulation of the theory.

A second concept from general relativity one might expect to be "fuzzed out" by quantization is the metric. There will not, presumably, be one specific metric of space-time, but some probability distribution of possible metrics. This "smearing out of the metric" might be expected to have significant physical consequences. For example, the divergences which arise in quantum field theory come about, at least in part, because integrals in momentum space extend to arbitrary large momenta. One uses "cutoffs" in momentum to obtain finite results. If the metric were "smeared out", one might expect this to result in natural cutoffs on such integrals. One might expect the divergence difficulties associated with quantum field theories to, at least, become less severe in the presence of a quantized metric. Furthermore, the singularities we have seen in general relativity might also be expected to disappear. The smoothing out from quantum theory could result in a smoothing over of these singularities. (Analogous phenomenon in atomic physics: Classically, an electron orbiting a point nucleus radiates, spirals inward, and eventually hits the nucleus. In quantum theory, this singularity disappears.)

The approaches to quantization of general relativity are normally based on analogies with quantum theories we understand: quantum electrodynamics, Schrödinger quantum mechanics for a particle, etc.

41. Linearized Approach to Quantization

Consider the linearized Einstein equations, (166) and (167). We have seen in Sect. 39 that this approach to general relativity results in a set of equations on a field γ_{ab} in flat space which bears a very close resemblance to Maxwell's equations of electrodynamics. The vector potential A_a for the electromagnetic field is replaced by the "potential for the gravitational field", γ_{ab}. These are both tensor fields in flat space-time. In some sense, the linearized Einstein equations represent an approximation to the full equations of general relativity.

The approach to quantization to be discusses in this section is based on the following idea. One regards γ_{ab} as just another classical field (on the same footing with, say, the electromagnetic vector potential). One attempts to use the conventional techniques of quantum field theory on this γ_{ab}. That is to say, one extends the analogy between electrodynamics and linearized general relativity to a quantization program for the latter. Using quantum electrodynamics as a model, one attempts to construct a "quantum gravidynamics".

In this section, we shall first summarize, in very broad and vague terms, the setting of quantum electrodynamics. We then remark that similar techniques could be applied to the linearized Einstein equation. Finally, we make some general comments on the resulting "quantum theory of gravitation".

There are two stages leading to quantum electrodynamics. In the first, one obtains the theory for free photons (the quantized version of the classical theory described by $\nabla^m \nabla_m A^a = 0$, $\nabla_a A^a = 0$). Next, one introduces interactions. For the free case, one proceeds, roughly, as follows. Consider the real (infinite-dimensional) vector space of (asymptotically well-behaved) solutions of Maxwell's equations with $J^a = 0$. One introduces on this vector space a suitable norm and a suitable complex structure. It thus becomes a Hilbert space H. This H represents the Hilbert space of one-photon states of the (source-free) Maxwell field. Next, one extends this description to states with many photons.

$$F = H^0 + H^1 + H^2 + H^3 + \cdots \tag{168}$$

where the superscripts denote "powers" of H. (More precisely, H^0 is the complexes, $H^1 = H$, H^2 is the tensor products of H with itself, H^3 the tensor product of H with H^2, etc. The sums are direct sums of Hilbert spaces.) This F, the Fock space, represents the states of the system (without sources). An element of H^n represents a state with n photons (and an element of H^0 a vacuum (zero photon) state). Thus, the general element of F consists of a linear combination of states with various numbers of photons. There are defined operators on F representing such things as "numbers of photons", "energy-momentum", etc. This, the theory of free (non-interacting) photons, is not very interesting, because nothing much happens.

One now introduces interactions. These are described by certain operators (on F, the Fock space for electrodynamics, and also on the Fock spaces for the other particles of interest, e.g., electrons). These interaction operators allow for the possibility of translations in which numbers of photons change (while, of course, numbers of other types of particles can also change). Thus, with the introduction of an interaction, one has the possibility of particle reactions' taking place. In this way, e.g., electron-photon scattering cross sections can be calculated, and comparison made with experiment.

Essentially the same program goes through, with little change, for the linearized Einstein equation. One introduces the Fock space of free gravitation states. For the interaction, the gravitons couple to the stress-energy of particles rather than (in the electromagnetic case) the charge. One calculates scattering processes, etc. involving gravitons. The proposal, then, is that one regard the result as representing a "quantization of general relativity".

There is certainty a sense in which the program above departs from the spirit of general relativity. One could, of course, criticize it on the grounds that it deals only with the linearized equations – not the full Einstein equation. This, however, is a deficiency only of our brief description – not of the program itself. One could just as well consider also the higher order terms in the perturbation expansion (i.e., in Sect. 39, one could take $\mathrm{d}^2/\,\mathrm{d}\lambda^2$, $\mathrm{d}^3/\,\mathrm{d}\lambda^3$, etc., at $\lambda = 0$, of Einstein's equation). These corrections would represent further possible interactions – they would be gravitation-gravitation interactions. Thus, one would regard the nonlinearity of Einstein's equation as allowing for the possibility of "gravitational field produced by gravitational field itself". Quantum-mechanically, gravitons create gravitons. The more terms included as interactions from the perturbation expansion, presumably, the closer the resulting field theory would approximate general relativity.

In my view, more serious objections are possible. A physical theory consists, of course, of more than merely the equations of that theory. In particular, general relativity consists of more than Einstein's equation. There is in addition to the equations, an overlay of concepts, attitudes, prejudices, etc. The concepts play at least as great a role in what the theory "is" as the equations. In general relativity, for example, there is the notion of assembling all possible events into a manifold. There is the notion of the metric on this manifold – an object with direct physical significance as giving the result of space and time measurements, and more indirect physical significance concerning gravitation. In short, general relativity is an integral part of what might be called the "space-time view of physics".

Where are these concepts from general relativity in the linearized version of quantized gravitation? One sees, at least, the rudiments of Einstein's equation, but, in my opinion, not the sense of the general theory of relativity. This is not to say, of course, that the linearized program is wrong. What it does seem to imply is that, if Nature behaves as described by this approach, then the general theory of relativity has been an unfortunate – and expensive in terms of time and effort – detour.

42. Canonical Approach to Quantization

As an alternative to the linearized approach, we now discuss the canonical approach to quantization. Perhaps this approach displays a greater respect for the integrity of general relativity than the linearized approach (and, for this reason, it is a good example of an alternative to the linearized). On the other hand, the canonical approach is, it seems to me, a bit on the simple-minded, naive side. It takes, at its model, elementary Schrödinger quantization of a particle. But quantum theory has advanced considerably since its beginnings. We begin with a brief review of Schrödinger quantization. We then attempt to carry over, as directly as possible, these ideas to general relativity. The result is an imprecise, but suggestive, program for obtaining a quantum theory of gravitation.

Consider a single particle. We can describe the particle by its position x and momentum p. Thus, as the particle moves around in time, the motion is described by $x(t)$ and $p(t)$. The dynamics of the particle are described by a pair of differential equations which express $\dot{x}$ and $\dot{p}$ as functions of x and p. Thus, if one specifies the values of x and p at some initial time, then the equations of motion determine x and p for all future times. The initial data for the particle consist of the values of x and p. It usually turns out that the equations of motion for the particle can be cast into the following form. One can find a certain function $H(x,p)$ of x and p such that

$$\dot{x} = \frac{\partial}{\partial p} H(x,p) \qquad \dot{p} = -\frac{\partial}{\partial x} H(x,p) \tag{169}$$

If the equations of motion can be cast into the form (169), they are said to be in *Hamiltonian form*. The function $H(p,q)$ is called the *Hamiltonian* of the system.

The Schrödinger quantization scheme is applicable to classical system whose equations of motion have been placed in Hamiltonian form. The initial data, x, p, are replaced by a single, complex-valued wave function, $\psi(x)$. Instead of $x(t), p(t)$, we have $\psi(x,t)$. Thus, the motion of the system in time is described via time-dependence in ψ. The Hamiltonian equations of motion, (169) are replaced by

$$-\frac{\hbar}{i}\frac{\partial}{\partial t}\,\psi = H\left(x, \frac{\hbar}{i}\frac{\partial}{\partial x}\right)\,\psi \tag{170}$$

where $H(x, \frac{\hbar}{i}\frac{\partial}{\partial x})$ means "replace p in $H(x,p)$ by the differential operator $\frac{\hbar}{i}\frac{\partial}{\partial x}$" (a rather vague prescription). Thus, given $\psi(x,t_0)$ for some value of t_0, (170) determines $\psi(x,t)$ for all t.

The Hilbert space for this quantum theory consists of the (complex) vector space of all (sufficiently well-behaved) solutions of (170). One then introduces position, momentum, energy operators, etc.

The idea is, firstly, to try to express the equations of general relativity in "Hamiltonian form". Then, one applies the Schrödinger prescription to obtain a quantum theory. Recall the initial-value formulation of general relativity. The initial data consist of the induced metric h_{ab} and the extrinsic curvature Π^{ab}. These evolve with time according to the equations

$$\begin{aligned} \dot{h}_{ab} &= 2\varphi\Pi_{ab} \\ \dot{\Pi}^{ab} &= -D^a D^b \varphi - 2\varphi \Pi^{am}\Pi^{b}{}_{m} - \varphi\Pi\Pi^{ab} + \varphi\mathscr{R}^{ab} \end{aligned} \tag{171}$$

where φ is the evolution function. One first task is to re-express (171) in Hamiltonian form.

Set $p^{ab} = \Pi^{ab} - \Pi h^{ab}$. This p^{ab} is, it turns out, more closely analogous to the p for a particle than Π^{ab}. Rewriting (171) in terms of p^{ab}, we obtain

$$\begin{aligned} \dot{h}_{ab} &= 2(p_{ab} - \frac{1}{2} p\, h_{ab}) \\ \dot{p}^{ab} &= -D^a D^b \varphi + \varphi \mathscr{R}^{ab} - 2\varphi\, p^a{}_m p^{bm} \\ &\quad + \frac{3}{2}\varphi\, p\, p^{ab} - 2\varphi\, h^{ab} p^{mn} p_{mn} + \frac{1}{4}\varphi\, p^2 h^{ab} \end{aligned} \tag{172}$$

We wish to express these equations in Hamiltonian form. This is in fact possible: choose for the Hamiltonian

$$H = -\int_S \varphi\, (\mathscr{R} - p^{mn} p_{mn} + \frac{1}{2} p^2)\, \mathrm{d}V \tag{173}$$

where the integral extends over the entire 3-manifold S. (We ignore questions of convergence of integrals. We shall also allow ourselves to throw away surface terms at will. Such details are unimportant at this stage of theory-building.) Note that $H(h_{ab}, p^{ab})$ does indeed assign a real number to each choice of data, $(h_{ab},\, p^{ab})$, as we would want.

Thus, we have a Hamiltonian formulation of the initial-value formulation of general relativity (ignoring for the moment, the question of constraints). Note that we go to the initial-value formulation of general relativity because, in the one-particle discussion, time played a special role. An analogy could be made to general relativity only reintroducing a "time" there. That is precisely what the initial-value formulation accomplishes. In a certain sense we have, already at this stage, violated the spirit of general relativity.

The next step is to write down the wave function. Instead of the $\psi(x)$ in the Schrödinger theory, we have $\psi(h_{ab})$, a complex-valued function of the collection of all positive-definite metrics on the (fixed) three-dimensional manifold S. We wish to permit evolution, so we write $\psi(h_{ab}, t)$. The Schrödinger equation, (170) becomes, using (173),

$$-\frac{\hbar}{i}\frac{\partial}{\partial t}\psi = \left(\frac{\hbar}{i}\right)^2 \left(h^{ac}h^{bd} - \frac{1}{2}h^{ab}h^{cd}\right) \frac{\delta^2 \psi}{\delta h_{ab}\, \delta h_{cd}} - \mathscr{R}\,\psi \tag{174}$$

where, $\delta/\delta h_{ab}$ refers (rather imprecisely) to functional derivatives. Thus, just carrying over the analogy with Schrödinger quantization, we are led to describe the "quantum gravitational field" by a complex-valued function, $\psi(h_{ab}, t)$, on the space of all positive-definite metrics on S, and on t, This function must satisfy the "gravitational Schrödinger equation", (174).

We have, up till now, ignored the constraint equations, (134) and (135). Clearly, one is doing something essentially wrong if he simply ignores certain equations: we have yet fully incorporated Einstein's equation into our theory. First note that, in terms of p^{ab}, the constraints take the form

$$D_b p^{ab} = 0 \tag{175}$$

$$\mathscr{R} - p^{ab} p_{ab} + \frac{1}{2} p^2 = 0 \tag{176}$$

The question is: How do we "incorporate" these equations into the theory? The most naive answer is simply to incorporate them as operator equations, using

the replacement of p_{ab} by $(\hbar/i)\,\delta/\delta h_{ab}$. Let's try this to see what happens. The classical constrain (175) would then be replaced by the following condition on our wave function

$$D_b\left(\frac{\hbar}{i}\frac{\delta\psi}{\delta h_{ab}}\right) = 0 \tag{177}$$

To interpret (177) multiply by an arbitrary vector field v_a on S, and integrate by parts to obtain

$$\int_S \left(\frac{\delta\psi}{\delta h_{ab}}\right)(D_{(a}v_{b)})\,\mathrm{d}V = 0 \tag{178}$$

The validity of (177) is equivalent to the validity of (178) for all v_a. But (178) is easy to interpret. Note that $D_{(a}v_{b)} = \frac{1}{2}\pounds_v h_{ab}$. Thus, $D_{(a}v_{b)}$ is (up to a factor) the rate of change of h_{ab} under the diffeomorphism generated by motions along the integral curves of v_a. Therefore (by the chain rule), (178) states that the rate of change of $\psi(h_{ab})$, as h_{ab} changes by the diffeomorphism generated by v_a, is zero. To say it another way, (178) requires that, if h_{ab} and h'_{ab} are two metrics on S which differ by a diffeomorphism on S (i.e., if h^{ab} and h'^{ab} are isometric), then $\psi(h_{ab}) = \psi(h'_{ab})$. We can write this symbolically as $\psi = \psi(\text{geometry})$. This conclusion is also reasonable physically. The physics of isometric metrics is identical (the only difference being the labeling of points of S). Thus, one might expect ψ to assume the same value on two such metrics. To summarize, the constraint equation (175) leads to the quantum condition (177) which, geometrically, means that ψ is invariant under replacing h_{ab} by the result of subjecting h_{ab} to a diffeomorphism in S.

We now repeat for (176). It is not hard to guess what the answer will be: (176) requires that $\psi(h_{ab}, t)$ be invariant under motions in time. That this is indeed the case can be seen immediately by noting that the Hamiltonian of our theory, (173), is just an integral of the constraint (176). Hence, the quantum version of (176) is precisely the condition that the right side of (174) vanishes. Thus, we require $(\partial/\partial t)\psi(h_{ab}, t) = 0$, i.e., that ψ in fact be independent of t. (It's a good thing. The interpretation of this "t" was always rather obscure, anyway.)

Thus, the constraints, (175) and (176), are related to the section of diffeomorphisms in space-time. This is expressed, in the quantum theory, by certain invariance of the wave function. In fact, we have seen these notions once before: in the gauge transformations in the linearized theory. These gauge transformations, again represented the action of diffeomorphisms. Thus, gauge (linearized version), action of diffeomorphisms (full theory), constraints (initial-value formulation), and invariance of wave function (quantum theory) all are manifestations of essentially the same thing.

We summarize by stating the formalism of this "theory". The Hilbert space is the space of all complex-valued functions $\psi(h_{ab})$ on the space of positive-definite metrics on S, such that ψ is invariant under the action (on h_{ab}) of diffeomorphisms on S, and such that

$$\left(\frac{\hbar}{i}\right)^2 \left(h^{ac}h^{bd} - h^{ab}h^{cd}\right)\frac{\delta^2\psi}{\delta h_{ab}\,\delta h_{cd}} - \mathscr{R}\,\psi = 0 \tag{179}$$

We put theory in quotation marks because we have here merely an equation and a few words. What does it all mean? What is the measurement situation? What would it be like to live in such a quantum space-time? What is the correspondence limit? To what extent have the principles of quantum theory been incorporated? To what extent is this theory a "fuzzing out" of general relativity?

What alternative formulations are available, and how do they compare with this one?

We are today a good way from satisfactory answers to questions of this sort.

Robert Geroch

Integration Theory

©1979 University of Chicago

1. Introduction

Three objects lie at the beginnings of a theory of integration: a base space X over which integrals are to be performed, a space M in which the measures ("areas") of subsets of X are to be valued, and a space G in which the functions on X to be integrated are valued. In an elementary textbook on integration, X is normally the set of real numbers, M the non-negative reals, and G again the reals. That is to say, one learns to integrate real-valued functions over the reals, using a non-negative real measure. More advanced textbooks deal with more general possibilities, typically X any set, M the reals, and G the complexes.

On the other hand, many of the applications of integration in both physics and mathematics require a more general notion of an integral. In physics, for example, it might be convenient to regard the momentum density of a system in Euclidean space as a measure on that space (so the "area" of a region of Euclidean space would be the total momentum of that part of the system in the given region). Then one would want to choose for M a three-dimensional vector space. If, under this set-up, one wished to evaluate the total angular momentum of the system as an integral, then one would like to choose for G also a three-dimensional vector space. If, to take a second example, one wished to treat the spectral theorem in a Hilbert space, then one would like to choose for M the set of bounded operators on the Hilbert space, and for G either the reals or the Hilbert space itself. It is not difficult to find other, similar examples.

Thus, one faces the situation that the standard mathematical treatment of integration seems to be too narrow for many of the natural applications of this subject. This is not to say that more general treatments are unavailable: they are, but normally one pays for any broadening of the framework of integration theory by having to read an entire book. There are at least three courses of action open to one who wishes to use integration in a context more general than that with which he is familiar: i) try to restate the definitions and reprove the theorems one needs for the new context, ii) try to so reformulate the question that one can get by with only a more elementary notion of an integral, or iii) try to guess, from one's

experience with integration, what will be true. The first can be time-consuming, the second mis-directing, and the third dangerous.

What one would like to do in response to this difficulty is just what one always tries to do in such situations in mathematics. One would like to formulate a version of integration theory which has two properties: i) the framework is sufficiently broad that one can feel reasonably secure that every "integration-like notion" that one meets in one's work will fit into it, and ii) the framework is sufficiently narrow that one can have within it a good spectrum of interesting and useful theorems. The problem, then, is one of compromise between competing goals. One would like to work hard once, and only once, to isolate integration theory in its "purest form", with the hope that life will be easier thereafter.

We here carry out an attempt at such a program.

2. Measurable Spaces

A measurable space consists of two things - i) a set X, and ii) a collection $\mathcal{M}$ of subsets of X (Subsets in this collection are called measurable (sub)sets.) - subject to the following three conditions:

1. The empty set ϕ is measurable.
2. If A is measurable, so is A^c, the complement of A in X.
3. If $A_1, A_2, \ldots$ is any countable collection of measurable sets, then $\cup A_i$, their union, is measurable.

The set X will ultimately be the set "over which integration is to be performed". The measurable subsets of X will be those subsets "for which 'area' can be given sensible meaning". Although it is perhaps not obvious that the three conditions above reflect one's intuitive notion of "having an area", they at least look not unreasonable. One might imagine, in particular, that it should be possible to so arrange matters that every subset of X will be measurable, whence one should so modify the definition above. We shall see later that one gets into trouble if he tries to do this, and, more generally, that the notion of a measurable space seems to be appropriate.

Several facts follow immediately from the definition. The first two conditions imply that X itself is measurable, and the last two that any intersection of a countable collection of measurable sets is measurable. From the first and third conditions, together with this last observation, we have that any intersection or union of a finite collection of measurable sets is measurable.

Example. Let X be any set, and let $\mathcal{M}$ consist of all subsets of X. The three conditions above are immediate, and so we obtain a measurable space.

Example. Letting X be any set, and $\mathcal{M}$ consist only of ϕ and X, we obtain a measurable space.

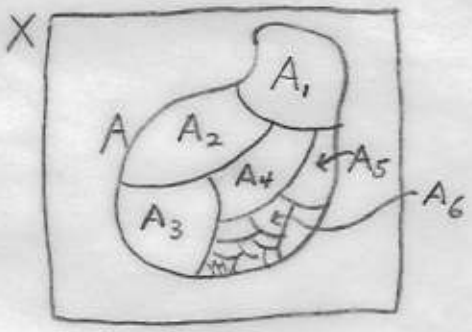
Example. Let X be any set, and let $\mathcal{M}$ consist of all subsets of X which are either empty, finite, or countable, or whose complement is either empty, finite, or countable. Conditions 1 and 2 are immediate, while condition 3 follows from: the union of any countable collection of sets, each either empty, finite, or countable, is empty, finite, or countable, while any union of sets, one of which has empty, finite, or countable complement, has empty, finite, or countable complement.

We shall shortly obtain some more interesting examples of measurable spaces.

There is one construction from which many useful measurable spaces arise. Let X be any set, and $\mathcal{A}$ any collection of subsets of X. Let $\mathcal{M}$ be the collection of all subsets of X which can be written as a countable union of countable intersections of sets, each of which is either the empty set, a set in $\mathcal{A}$, or the complement of a set in $\mathcal{A}$. Then X, $\mathcal{M}$ is a measurable space, called the measurable space <u>generated by</u> $\mathcal{A}$. Clearly, this $\mathcal{M}$ is "the smallest one which includes all the sets in $\mathcal{A}$, and which gives a measurable space". That is to say, one adds to $\mathcal{A}$ as few sets as one must to get a measurable space.

<u>Example</u>. Let X be any set, and let $\mathcal{A}$ consist of all subsets of X having but a single element. The measurable space generated by $\mathcal{A}$ is precisely the last example on the previous page.

It is convenient to introduce at this point one additional notion that we shall need later. Let X, $\mathcal{M}$ be a measurable space, and let A be any measurable set. A <u>partition</u> of A consists of a countable collection $A_1, A_2, \ldots$ of measurable sets, such that no two of these intersect, and such that their union is A. That is, a partition of A divides A into "a countable number of disjoint, measurable pieces". (Some A_i could be empty, so a partition could be effectively a division into a finite number of "pieces".)

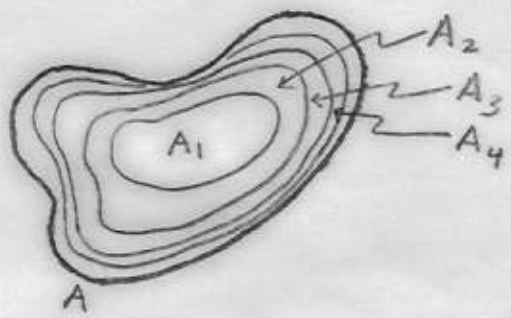


Given two partitions, $A_1, A_2, \ldots$ and $A'_1, A'_2, \ldots$, of A, we say that the first is <u>finer</u> than the second if every A_i is a subset of some A'_j (i.e., if the first has "smaller pieces"). We note the following fact: if $A_1, A_2, \ldots$ and $A'_1, A'_2, \ldots$ are partitions of A, then there exists a partition finer than each of these. [Proof: Consider the partition $A''_1, A''_2, \ldots$ whose sets are those of the form $A_i \cap A'_j$, for i,j = 1,2,3,...]

<u>Example</u>. Every partition is finer than itself. For X uncountable in the last example on page 3, X itself has no finest partition.

3. Topological Groups

We now have the notion of a set, together with those subsets of that set to which we shall assign area. What sort of object should we now select for "the space of all values the area might assume"? One could imagine attacking this question by trying to decide what the word "area" means to him. Fortunately, it turns out that there is a way to deal with such questions which is at the same time more specific and less conservative: one asks instead what operations one would sensibly expect to be able to do with areas. It seems to me that there are two such operations. First, one should be able to add and subtract areas. For example, one might wish to ask whether or not the area of the union of two disjoint sets is the sum of the areas of these sets. Second, one should be able to take limits of areas. For example, suppose one has a collection of sets, A_1, A_2, ..., each of which is a subset of its successor, and whose union is A. One would like to be able to ask whether or not the limit of the areas of the A_i is the area of A. Thus, the space of areas should be endowed with structure which permits sums and limits. It turns out, fortunately, that there is a standard mathematical object which has precisely this structure. We introduce it in the present section.



An abelian group consists of two things - i) a set G, and ii) a rule which assigns, to any two elements g and g' of G, a third element of G (written g + g', and called the sum of g and g') - subject to the following four conditions:

1. For any elements g, g', and g'' of G, (g + g') + g'' = g + (g' + g'').

2. For any elements g and g' of G, g + g' = g' + g.

3. There exists an element of G, written 0, such that g + 0 = g for every g in G.

4. For every g in G, there exists an element of G, written - g, such that g + (- g) = 0.

It is immediate from the first two conditions that any sum of a finite number of elements of G can be grouped in any way with parentheses, and added in any order, with the same result. From the last two conditions, $-0 = 0$.

Example. Let G be the real numbers, or the rational numbers, or the integers, or the complex numbers, with the sum addition of numbers in each case. One obtains, in each case, an abelian group.

Example. Let G be $\mathbb{R}^n$, the set of n-tuples $(r_1,\ldots,r_n)$ of real numbers. Let addition be $(r_1,\ldots,r_n) + (r'_1,\ldots,r'_n) = (r_1+r'_1,\ldots, r_n + r'_n)$. This is an abelian group, where $0 = (0,\ldots,0)$, and $-(r_1,\ldots,r_n) = (-r_1,\ldots,-r_n)$.

Example. The set of all real-valued functions on any set S, where addition is point-wise addition of functions, is an abelian group.

The four conditions for an abelian group are to represent the conditions one associates with the abstract operation of "addition". [The adjective "abelian" stands for the fact that condition 2 has been included in the list.] One can try to convince himself directly that the operations of "addition and subtraction of areas" would not be very useful without these four conditions.

It is convenient to introduce the following notational conventions. Instead of $g + (-g')$, we often write $g - g'$; for g in G and U a subset of G, we write $g + U$ for the subset of G consisting of all elements of G which can be written in the form $g + u$ with u in U; for U and U' subsets of G, we write $U + U'$ for the subset of G consisting of all elements of G which can be written in the form $u + u'$ with u in U and u' in U'; for U a subset of G, we write $-U$ for the subset of G consisting of all elements of the form $-u$ with u in U; for U and U' subsets of G, we write $U - U'$ for $U + (-U')$. Thus, for example: $U + U' = U' + U$, $U + (U' + U'') = (U + U') + U''$, $U + 0 = U$, $-(-U) = U$, etc. [Caution: It is false in general that $U - U$ contains only 0.]

We now have objects in which addition is available. We must next add to such objects structure which allows the taking of limits.

An abelian topological group consists of two things - i) an abelian group G, and ii) a collection $\mathcal{U}$ of subsets of G (Subsets in this collection are called 0-neighborhoods.) - subject to the following

five conditions:

1. The intersection of all 0-neighborhoods contains only 0.

2. The intersection of any two 0-neighborhoods is another.

3. Any subset of G which contains a 0-neighborhood as a subset is itself a 0-neighborhood.

4. For U any 0-neighborhood, so is - U.

5. For any 0-neighborhood U, there is a 0-neighborhood V such that U contains V + V as a subset.

One thinks of a 0-neighborhood as representing a certain "degree of closeness" to the element 0 of G, namely, for U a 0-neighborhood, an element g of G is close to 0 "by U's standards" if g is in U. The first condition requires that 0 be in every 0-neighborhood ("0 is close to 0 by anyone's standards"), and that every nonzero element of G fail to be in some 0-neighborhood. The second and third conditions amplify this viewpoint. The intuitive statement of the fifth condition is "you tell me how close (U) you want to be to zero, and I'll tell you how close (V) two elements of the abelian group must be to zero in order that their sum be within U of zero". Similarly for the fourth.

Example. Let G be the abelian group of real numbers. Call a subset U of G a 0-neighborhood if there exists a positive number ϵ such that every number r with $|r| < \epsilon$ is in U. Conditions 1, 3, and 4 are immediate. For condition 2, note that, if all r with $|r| < \epsilon$ are in U, and all r with $|r| < \epsilon'$ are in U', then all r with $|r| < \epsilon''$ are in $U \cap U'$, where ϵ'' is the smaller of ϵ and ϵ'. For condition 5, note that, if all r with $|r| < \epsilon$ are in U, then, letting V consist of all r with $|r| < \epsilon/2$, we have r' + r'' in U whenever r' and r'' are in V. Thus, we have an abelian topological group.

Example. Let G be the abelian group R^n. Call a subset U of G a 0-neighborhood if there exists a positive number ϵ such that $(r_1,\ldots,r_n)$ is in U whenever $[(r_1)^2 + \ldots + (r_n)^2]^{1/2} < \epsilon$. By essentially the same remarks as above, this G is an abelian topological group.

Example. Let G be any abelian group, and let the 0-neighborhoods consist of all subsets of G containing zero. The five conditions for an abelian topological group are immediate.

Let G be an abelian topological group. One might expect that, since we know what it means to be "close to 0" in G, one should also know what it means to be "close to any given element of G". For g in G, a g-neighborhood is a subset of G of the form g + U, where U is a 0-neighborhood. Thus, for example: g is in every g-neighborhood; g' is in g-neighborhood g + U when and only when g' - g is in U.

We shall often make use of the following generalization of condition 5: given any positive integer n, and any 0-neighborhood U, there is a 0-neighborhood V such that V + V + . . . + V (n times) is a subset of U. [Proof: By condition 5, choose 0-neighborhood V_1 with $V_1 + V_1$ a subset of U. By condition 5, choose 0-neighborhood V_2 with $V_2 + V_1$ a subset of V_1. Then $V_2 + V_2 + V_2 + V_2$ is a subset of U, etc.] The above property also holds if some of the "+'s" are replaced by "-'s". [Proof: First find 0-neighborhood V with V + V + ... + V (n times) a subset of U. Set $V' = V \cap (-V)$, a 0-neighborhood by conditions 2 and 4. Then V' + V' + . . . - V' (n times, with any combination of "+'s" and "-'s") is a subset of U.]

We originally introduced the notion of an abelian topological group in order to be able to take limits. Limits are now defined as follows. Let G be an abelian topological group, and let g_1, g_2, . . . be any sequence of elements of G. We say this sequence converges to element g of G if the following condition obtains: for any g-neighborhood g + U, there is an integer N such that g_n is in g + U whenever $n \geq N$. Intuitively, g_1, g_2, . . . converges to g if "whenever you tell me how close you want to be to g, I can tell you how far you must go along the sequence until every element thereafter will be that close to g". It is easily checked that this notion of convergence coincides, e.g., in the case of the abelian topological group of reals, with ordinary convergence of a sequence of real numbers. When sequence g_1, g_2, . . . converges to g, we say that g is the limit of the sequence. As an example of these notions, we prove the following: if sequence g_1, g_2, . . . converges to both g and g', then g = g'. Proof: Suppose $g \neq g'$. Choose, by condition 1, 0-neighborhood U with g - g' not in U. Let V be a 0-neighborhood with V - V a subset of U. Then there is an N with g_n in g + V for all $n \geq N$, and an N' with g_n in g' + V for all $n \geq N'$. Hence, there is some g_n in both g + V and g' + V, i.e., some g_n with both g_n - g and

$g_n - g'$ in V. Hence, $g - g' = (g_n - g') - (g_n - g)$ is in V - V, whence $g - g'$ must be in U. This contradiction establishes $g = g'$.

A sequence $g_1, g_2, \ldots$ in abelian topological group G is said to be a Cauchy sequence if the following condition obtains: for every 0-neighborhood U, there is an integer N such that $g_n - g_{n'}$ is in U whenever both $n \geq N$ and $n' \geq N$. Thus, a sequence is a Cauchy sequence if its elements "eventually become and remain as close to each other as one wishes". By contrast, a sequence converges to g if its elements "eventually become and remain as close to g as one wishes". Note in particular that convergence requires an element g of G to converge to, while "being Cauchy" is a property of sequences themselves. These intuitive characterizations suggest that a sequence which converges to some element of G must necessarily be a Cauchy sequence. In fact, this is true. Proof: Let $g_1, g_2, \ldots$ converge to g in G. Fix a 0-neighborhood U: we must find an N such that $g_n - g_{n'}$ is in U whenever $n \geq N$ and $n' \geq N$. Choose a 0-neighborhood V with V - V a subset of U. By convergence, there exists an N such that $g_n - g$ is in V whenever $n \geq N$. Hence, for $n \geq N$ and $n' \geq N$, both $g_n - g$ and $g_{n'} - g$ are in V, whence $g_n - g_{n'}$ is in V - V, whence $g_n - g_{n'}$ is in U. Thus, the sequence is Cauchy.

Example. Let G be the abelian topological group of reals, and consider the sequence $g_1 = 1/2$, $g_2 = 3/4$, $g_3 = 7/8$, . . . This is a Cauchy sequence, with limit $g = 1$. The sequence $g_1 = 1$, $g_2 = -1$, $g_3 = 1$, . . . is not Cauchy, and has no limit.

Example. Let G be the abelian topological group of rationals. Let $g_1 = 3$, $g_2 = 3.1$, $g_3 = 3.14$, . . . , $g_6 = 3.14159$, . . . (decimal expansion of pi). This is a Cauchy sequence. However, it does not converge to any element of G. [The only candidate for its limit would be pi, but pi is not an element of the abelian topological group G.]

Thus, a sequence which converges is always a Cauchy sequence, while, as the last example above shows, a sequence can be Cauchy without having a limit. An abelian topological group is said to be complete if every Cauchy sequence in that group converges to some element of the group. Thus, the abelian topological group of rationals is not complete. That the abelian topological group of reals is complete is one of the basic facts about the real number system.

4. Series

In this section, we introduce the notion of an "infinite sum" in a complete abelian topological group.

Let G be a complete abelian topological group. Denote by G* the set whose elements are those of G, together with one additional abstract element, written "∞", and said to be infinite. Those elements of G* which are also in G are said to be finite. We extend addition and subtraction from G to G* as follows: for g and g' in G*, let g + g' be g + g' (in G) if g and g' are finite, and ∞ if either or both are infinite; for g in G*, let - g be - g (in G) if g is finite, and ∞ if g is infinite. Note that this G*, with this definition of addition, is not even an abelian group (for, while conditions 1, 2, and 3 for an abelian group are indeed satisfied, the fourth is not: $\infty + (-\infty) = \infty$). The purpose of this enlargement of G is to allow us to consider cases when limits do and do not exist simultaneously, i.e., to avoid having to make statements and proofs which involve numerous particular cases.

Let G be a complete abelian topological group. A sequence $g_1, g_2, \ldots$ of elements of G* is called a series. For such a series, its sum, written Σg_i, is defined as follows:

1. $\Sigma g_i = g$, where g in G* is finite, if, for every g-neighborhood g + U in G, there is a finite collection K of positive integers such that, for every finite collection K' containing K as a subset, $\sum_{K'} g_i$ is in g + U.

2. $\Sigma g_i = \infty$ otherwise.

Note that this sum is unique (i.e., if both g and g' satisfy the first condition, then g = g'), and that the sum is defined for any series. Note also that, if any g_i is infinite, then Σg_i is also infinite. A series is said to converge if its sum is finite.

Example. Let G be the complete abelian topological group of reals, and consider the series $g_1 = 1/2$, $g_2 = 1/4$, $g_3 = 1/8$, . . . Then $\Sigma g_i = 1$.

Example. Let G be the complete abelian topological group of reals, and consider the series $g_1 = 1/2$, $g_2 = -1/3$, $g_3 = 1/4$, $g_4 = -1/5$, . . . Then $\Sigma g_i = \infty$, for, given any finite collection K of positive integers, there is a finite collection K' containing K as a subset with $\sum_{K'} g_i$ as large as one wishes.

These two examples suggest that convergence of a series in a complete abelian topological group is analogous to absolute convergence of a series of real numbers. This analogy is further supported by at least two facts. First, as is easily checked, a series in the complete abelian topological group of reals converges if and only if that series converges absolutely in the usual sense. Second, the sum of a series is independent of the order of the $g_1, g_2, \ldots$ [Proof: The definition of the sum doesn't refer to the order.] It is this last property which makes the notion of the sum of a series useful for our applications. We shall want to take the sum of the "areas" of an infinite collection of disjoint sets. The sum as defined above would be inappropriate were it not independent of order. Note finally that this generalization of "absolute convergence" from the reals to a general complete abelian topological group is a somewhat subtle business: that one can take the absolute value of a real number is needed for the notion of absolute convergence, while there is nothing like the "absolute value" of an element of a complete abelian topological group.

There are a number of properties of sums of series of which we shall make use. We next introduce these properties.

Let G be a complete abelian topological group, and let $g_1, g_2, \ldots$ be a series in G*. A series $g'_1, g'_2, \ldots$ is said to be a subseries of this series if there exists a one-to-one mapping λ from the set of positive integers to itself such that $g'_i = g_{\lambda(i)}$ for $i = 1,2,\ldots$ [Recall that a mapping λ is said to be one-to-one if, whenever $i \neq j$, $\lambda(i) \neq \lambda(j)$.] Thus, a subseries is one that contains only "some of the terms" (namely, those g_j for j of the form $\lambda(i)$) of the original series. We have

Theorem 1. In a complete abelian topological group, every subseries of a convergent series converges.

Proof: Let the series $g_1, g_2, \ldots$ converge, $\Sigma g_i = g$, and let $g'_1, g'_2, \ldots$ be a subseries. For U any 0-neighborhood, there is a finite set K of positive integers with $\sum_{K'} g_i$ in $g + U$ whenever K' contains K as a subset. Hence, for K' and K'' such, $\sum_{K'} g_i - \sum_{K''} g_i$ is in $U - U$. We conclude: any sum of a finite collection of g_i, including no g_i with i in K, is in $U - U$.

Now let $h_1, h_2, \ldots$ be the sequence with $h_n = \sum_{i}^{n} g'_i$. Let N be the largest integer with $\lambda(N)$ in K. Then, for $n \geq N$ and $n' \geq N$, $h_n - h_{n'}$ is in U,- U, for this difference is a sum of a finite collection of g_i including none with i in K. We conclude that the sequence $h_1, h_2, \ldots$ is a Cauchy sequence. Since G is complete, this sequence converges to some finite g' in G*.

The proof is completed by showing $\Sigma g'_i = g'$. Fix 0-neighborhood U, and let N be as in the previous paragraph. Choose integer $\bar{N} \geq N$ with $h_{\bar{N}} - g'$ in U (possible, since $h_1, h_2, \ldots$ converges to g'). Let L consist of the integers from 1 to $\bar{N}$. Then, if L' contains L as a subset, we have $\sum_{L'} g'_i = \sum_{L} g'_i + \sum_{L'-L} g'_i = h_{\bar{N}} + \sum_{L'-L} g'_i$. But the first term on the right is in g' + U, while the second, as a sum of g_i including none with i in K, is in U - U. Hence, $\sum_{L'} g'_i$ is in g' + U + U - U. We conclude that the sum of the series $g'_1, g'_2, \ldots$ is g'. /

Example. That the series in the last example on page 10 does not converge is immediate from Theorem 1 and the fact that
1/2 + 1/4 + 1/6 + 1/8 + . . . = .

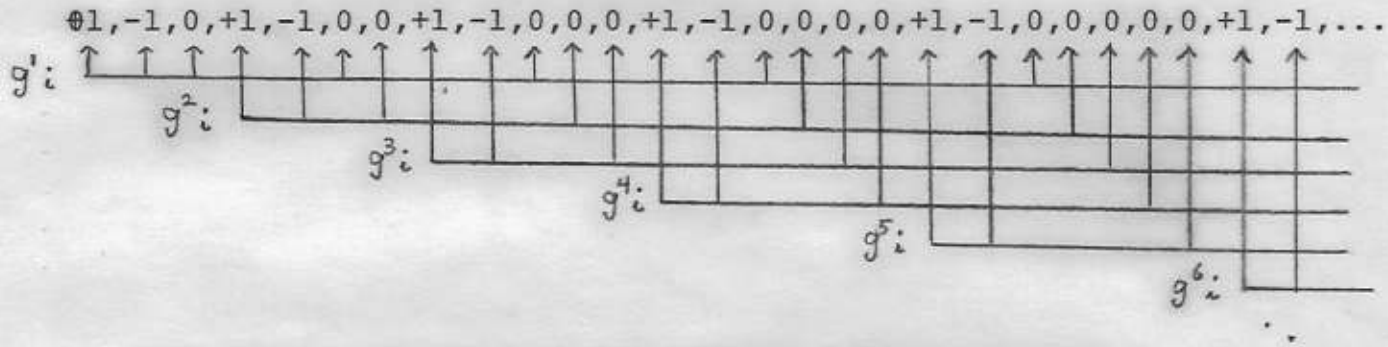
Note that we used completeness of the abelian topological group G in the proof of Theorem 1. In fact, it is not difficult, using the idea of the last example of page 9, to show that Theorem 1 would be false without completeness. It was essentially in order to obtain Theorem 1 that we originally introduced completeness, and it will be in order to use this theorem that we shall always assume it.

Let $g_1, g_2, \ldots$ be a series in G*, where G is a complete abelian topological group. The n series $g^1_1, g^1_2, \ldots,$ $g^2_1, g^2_2, \ldots, \ldots, g^n_1, g^n_2, \ldots$ are said to be a decomposition of this series if there are n mappings λ^j (j = 1,2,...,n) from the set of positive integers to itself such that i) for each j = 1,...,n and i = 1,2,..., $g^j_i = g_{\lambda^j(i)}$, and ii) for each positive integer k, there is one and only one pair (i,j) such that $k = \lambda^j(i)$. A decomposition of a series into an infinite number of series is defined similarly. Thus, a decomposition of a series consists of a collection of subseries such that every element of the original series is contained in one and only one of the subseries. A decomposition represents a "regrouping of all the terms of the original series into various subseries".

Now let g_i be a series in G*, and let g^j_i be a decomposition of this series. We are interested in the following question: when does

$\sum_j \sum_i g^j{}_i = \Sigma g_i$? That is to say, when can we first sum each subseries, and then take the sum of the results, getting the same answer as that of summing the original series? It is easy to think of examples in which these two procedures yield the same result. We give an example in which they do not.

Example. Let G be the complete abelian topological group of reals. Let g_i be the series with successive terms
+1, -1, 0, +1, -1, 0, 0, +1, -1, 0, 0, 0, +1, -1, 0, 0, 0, 0, +1, ...
Consider the decomposition indicated below:



Thus, each of the subseries in this decomposition has successive terms +1, -1, 0, 0, . . . Hence, $\sum_i g^j{}_i = 0$ for each j, whence $\sum_j \sum_i g^j{}_i = 0 + 0 + \ldots = 0$. On the other hand, $\Sigma g_i = \infty$.
Examination of this example shows that its mechanism seems to require two things: i) the original series fails to converge, and ii) the decomposition is into an infinite number of subseries. We now wish to generalize these observations to theorems about series in a general complete abelian topological group.

Theorem 2. Let g_i be a series in G*, where G is a complete abelian topological group. Let $g^j{}_i$ (j = 1,...,n, i = 1,2,...) be a decomposition of this series into a finite number of subseries. Then $\sum_j \sum_i g^j{}_i = \Sigma g_i$.
Proof: For j = 1,...,n, set $h^j = \sum_i g^j{}_i$. First note that, if any h^j is infinite, then, by Theorem 1, the conclusion of the present theorem holds, for each side of the concluding equation is ∞. Hence, we may assume that each h^j is finite. Fix a 0-neighborhood U. Then, for each j, there is a finite set K^j of positive integers such that, for any finite K'^j containing K^j as a subset, $\sum_{K'j} g^j{}_i$ is in $h^j + U$. Let K be the set of all integers of the form $\lambda^j(i)$ with i

in K^j. Then, for any finite K' containing K as a subset, $\sum_{K'} g_i$ is in $h^1 + \ldots + h^n + U + \ldots + U$ (with n U's). But this is just the statement that $\sum g_i = h^1 + \ldots + h^n$. /

Thus, for a decomposition into a finite number of subseries, one can always first sum the subseries, and then add, to obtain the sum of the original series. It is of interest to note that Theorem 1 is an immediate special case of Theorem 2 (although we could not proceed in this way, for we used Theorem 1 in the proof of Theorem 2).

The second case is that of a decomposition of a convergent series.

Theorem 3. Let g_i be a convergent series in G*, where G is a complete abelian topological group. Let $g^j{}_i$ (i,j = 1,2,...) be a decomposition of this series into subseries. Then $\sum_j \sum_i g^j{}_i = \sum g_i$.

Proof: For each j, set $h^j = \sum_i g^j{}_i$, so, by Theorem 1, each h^j is finite. We must show that $\sum_j h^j = g$, where $g = \sum g_i$. Let U be a 0-neighborhood, and let K be a finite set of positive integers such that, for any finite K' containing K as a subset, $\sum_{K'} g_i$ is in g + U. Let L be the (finite) set of j's such that $\lambda^j(i)$ is in K for some i. Let finite L' contain L as a subset. Then, by Theorem 2, $\sum_{L'} h^j = \sum_{\bar{K}} g_i$, where $\bar{K}$ consists of all integers of the form $\lambda^j(i)$ with j in L'. But $\bar{K}$ contains K as a subset, and so $\sum_K g_i$ is in g + U + U. Hence, $\sum_{L'} h^j$ is in g + U + U. Thus, we have shown that, for every 0-neighborhood U, there is a finite collection L of positive integers such that, for every finite L' containing L as a subset, $\sum_{L'} h^j$ is in g + U + U. That is to say, we have shown $\sum_j h^j = g$. /

Thus, for any decomposition of a convergent series, one can add the sums of the subseries to obtain the sum of the original series. The example on the previous page shows that we can hardly expect to do better than Theorems 2 and 3. The only case when $\sum_j \sum_i g^j{}_i \neq \sum g_i$ is when the range of j's is infinite, the left side is finite, and the right side infinite.

Finally, we remark that, for the case of the complete abelian topological group of reals, Theorems 1, 2, and 3 are standard facts about sums of series.

5. Measure Spaces

We have now completed the preliminaries. We have introduced the notion of a measure space (with partitions and their properties), together with that of a complete abelian topological group (with series and their properties). What we now wish to do is put these together. There results the central mathematical object of this subject.

A measure space consists of three things - i) a measurable space $X, \mathcal{M}$, ii) a complete abelian topological group M, and iii) a mapping μ from the set $\mathcal{M}$ to the set M* (For A a measurable set, the element $\mu(A)$ of M* is called the measure of A.) - subject to the following two conditions:

1. The measure of the empty set is zero.
2. Every measurable set A has a partition such that, for any refinement $A_1, A_2, \ldots$ of that partition, $\Sigma \mu(A_i) = \mu(A)$.

Example. Let X be the set of positive integers, and let every subset of X be measurable. Let M be any complete abelian topological group, and fix elements $m_1, m_2, \ldots$ of M*. Let $\mu(\emptyset) = 0$, and, for any nonempty subset A of X, let $\mu(A) = \sum_A m_i$. This is a measure space. The first condition is immediate. For the second condition, note that every A has a finest partition, namely, that for which each A_i is a single point of A. This partition suffices for the second condition.

Thus, a measure space consists of a measurable space, together with an assignment of an "area" (in M*) to each measurable set (that's what μ does), such that "areas are additive" in the sense of condition 2. Why is condition 2 what it is? Why does one not simply require $\Sigma \mu(A_i) = \mu(A)$ for every partition? Consider the example above, with M the reals, and with $m_1 = +1$, $m_2 = -1$, $m_3 = +1$, etc. Let A = X, and let A_1 consist of elements "1" and "2", A_2 consist of elements "3" and "4", etc. Then $A_1, A_2, \ldots$ is a partition of A. But $\Sigma \mu(A_i) = 0 + 0 + \ldots = 0$, while $\mu(A) = \infty$. Such examples are of practical interest, and so we would not like to rule them out in our definition of a measure space. Thus, condition 2 only requires that "areas are additive when the set A is cut into sufficiently small pieces". Note in particular that

the second condition implies that any partition of measurable A has a refinement with $\Sigma\mu(A_i) = \mu(A)$. What about the first condition? Should it not follow somehow from the second? Let $X, \mathcal{M}$ be any measurable space, and let $\mu(A) = \infty$ for any measurable A, including the empty set. Then condition 2 is satisfied, but 1 is not. We shall see shortly that this, uninteresting, example is the only one that the first condition rules out. We shall often say "... set A with measure ..." to mean "... measurable set A with measure ...". A partition of measurable A will be said to be <u>fine</u> if $\Sigma \mu(A_i) = \mu(A)$ for every refinement of it.

Three basic facts about measure spaces are direct consequences of the three facts about series in the previous section.

<u>Theorem</u> 4. Every measurable subset of a set with finite measure has finite measure.

Proof: Let A have finite measure, and let B be a measurable subset of A. Choose fine partition $B_1, B_2, \ldots$ of B. Choose a fine partition of A finer than the partition $A, -B, B_1, B_2, \ldots$ For this partition, $\mu(A) = \Sigma \mu(A_i)$, while $\mu(B) = \sum_K \mu(A_i)$, where K consists of those i with A_i a subset of B. Hence, $\mu(A)$ is the sum of a convergent series, while $\mu(B)$ is the sum of a subseries. By Theorem 1, $\mu(B)$ is finite. /

<u>Theorem</u> 5. Let $A_1, A_2, \ldots, A_n$ be disjoint measurable sets with union A. Then $\mu(A) = \Sigma\mu(A_i)$.

Proof: Choose a fine partition of A, and a fine partition of each A_i. Taking intersections of these subsets of A, we obtain a partition $A'_1, A'_2, \ldots$ of A. For this partition, $\mu(A) = \Sigma \mu(A'_i)$, each $\mu(A_i)$ is the sum of a subseries of this series, and these subseries form a decomposition of the original series. By Theorem 2, $\mu(A) = \Sigma \mu(A_i)$. /

Note that we did not use the first condition in the proof of Theorem 5, and that, from Theorem 5, $\mu(A) = \mu(A) + \mu(\phi)$ for any A. Thus, the first condition follows from the second unless every set has infinite measure.

<u>Theorem</u> 6. Let $A_1, A_2, \ldots$ be a partition of A with $\mu(A)$ finite. Then $\mu(A) = \Sigma \mu(A_i)$.

Proof: Choose a fine partition of A which, restricted to each A_i, is also fine. The result then follows from Theorem 3. /

These three theorems show that "area is additive whenever it should be", i.e., except for the situation illustrated in the example at the bottom of page 15. Note that we don't have to put these natural properties in "by hand": they follow already from the single condition 2, together with the properties of complete abelian topological groups (properties over which we, of course, have no control).

We summarize with a table:

When Does $\sum \mu(A_i) = \mu(A)$?

	finite number of A_i	infinite number of A_i	a still finer partition
$\mu(A)$ finite	yes	yes	yes
$\mu(A)$ infinite	yes	not necessarily	if fine enough

6. Example: Lebesque Measure

A single example dominates this subject. It provides the intuitive background for measure theory, it motivates and illustrates the definitions, and, in fact, it is the special case of interest for many applications. We now introduce this example.

To specify a measure space, one must say what the set X is, what the measurable subsets are (such that the three conditions for a measurable space are satisfied), what the complete abelian topological group M is, and what measureiis to be assigned to each measurable set (such that the two conditions for a measure space are satisfied). For Lebesque measure, X will be the set of reals, and the complete abelian topological group M will be that of the reals. We postpone, for a moment, the statement of what the measurable sets are to be, and what their measures are to be.

The idea is the following. Consider an open interval (a,b) in the reals (i.e., $a \leq b$, and (a,b) is the set of all real numbers r with $a < r < b$). One would like to assign to this interval measure $b - a$, the "length" of the interval. I suppose that one would, if he could, stop right here: let the measurable sets be the open intervals, and let the measure of such an interval be its length. Unfortunately, one cannot stop here, for, e.g., the open intervals in the reals do not even form a measurable space. (Neither the complement of an open interval, nor the union of a countable collection of open intervals, gives in general an open interval.) Apparently, the best hope of eventually getting a measure space, starting from "lengths of open intervals" would be to somehow enlarge significantly the collection of sets which will be called measurable. In fact, we have already a mechanism for doing just such a thing: take the measurable space on the reals generated by the open intervals. That is to say, one might take for the measurable sets those which can be written as a countable union of countable intersections of open intervals and their complements. In this way, we would indeed arrive at a collection of "measurable sets", but we would at the same time introduce a new difficulty: what are the measures of these sets to be? The simple notion of "length", which works for the open intervals, would need to undergo a significant generalization in order that we may assign "length" to these very complicated sets. It should not be

suprising that a more fruitful course is a more oblique one: instead of worrying first about what the measurable sets are to be, and then about what their measures are to be, we attempt to deal with both questions simultaneously.

Let A be any subset of the reals. A covering of A (by open intervals), consists of a countable collection, $I_1 = (a_1,b_1)$, $I_2 = (a_2,b_2)$, . . . , of open intervals such that A is a subset of their union. For any such covering, $\mathcal{C}$, denote by $L(\mathcal{C})$ the (possibly infinite) element $L(\mathcal{C}) = (b_1 - a_1) + (b_2 - a_2) + \ldots$ of $\mathbb{R}^*$. That is to say, $L(\mathcal{C})$ is the sum of the lengths of the intervals which make up $\mathcal{C}$. Of course, the intervals could overlap, or they could cover whole regions of $\mathbb{R}$ not included in A, and so $L(\mathcal{C})$ could be large although A is a rather "small" set. To get an element of $\mathbb{R}^*$ more characteristic of A, we proceed as follows: let $\bar{\mu}(A)$, the outer measure of A, be the greatest lower bound of the set of $L(\mathcal{C})$'s, for $\mathcal{C}$ ranging over coverings of A.

Example. Let A be the open interval (0,1). Consider the covering $\mathcal{C}$ given by $I_1 = (0,1)$, $I_2 = (0,0)$, $I_3 = (0,0)$, . . . Then $L(\mathcal{C}) = 1 + 0 + 0 + \ldots = 1$. Clearly, this is the smallest $L(\mathcal{C})$ one can get for $\mathcal{C}$ a covering of A. Hence, $\bar{\mu}(A) = 1$.

Example. $\bar{\mu}(\mathbb{R}) = \infty$.

Example. Let A consist of the set of all rationals, ordered in some way: $r_1, r_2, \ldots$ Fix any positive number ϵ . Consider the covering $\mathcal{C}$ of A given by $I_1 = (r_1 - \epsilon , r_1 + \epsilon)$, $I_2 = (r_2 - \epsilon/2, r_2 + \epsilon/2)$, $I_3 = (r_3 - \epsilon/4, r_3 + \epsilon/4)$, . . . Then $L(\mathcal{C}) = 2\epsilon + 2\,\epsilon/2 + 2\,\epsilon/4 + \ldots = 4\epsilon$. Since $\bar{\mu}(A)$ is the greatest lower bound of all such $L(\mathcal{C})$'s, we must, in particular, have $\bar{\mu}(A) \leq 4\epsilon$. Since this must be true for every positive ϵ, we have $\bar{\mu}(A) = 0$.

Thus, "outer measure" seems to be an appropriate generalization of "length" to all subsets of the reals. One might imagine that we are now nearly done: let the measurable sets be all subsets of the reals, and let the measure of any such subset be its outer measure. We certainly have a measurable space (first example on page 3), and so all that remains is to check the two conditions for a measure space. The first condition is of course immediate - but it turns out that the second one fails. (We shall give an example shortly.) What we must do, to get a measure space, is allow only certain subsets

of the reals to be called measurable.

We first note that outer measure has the following three properties:

1. $\bar{\mu}(\phi) = 0$.

2. For A a subset of B, $\bar{\mu}(A) \leq \bar{\mu}(B)$.

3. For $A_1, A_2, \ldots$ any subsets of the reals, $\bar{\mu}(\cup A_i) \leq \Sigma \bar{\mu}(A_i)$. (where, of course, every element of R^* is $\leq \infty$)

Proof:

1. Immediate.

2. Fix any positive number ϵ. Since $\bar{\mu}(B)$ is the greatest lower bound of $L(\mathcal{C})$'s with $\mathcal{C}$ a covering of B, there exists such a covering, $\mathcal{C}$, such that $L(\mathcal{C}) \leq \bar{\mu}(B) + \epsilon$. But, since A is a subset of B, this $\mathcal{C}$ is also a covering of A, whence $\bar{\mu}(A) \leq L(\mathcal{C})$. Hence, $\bar{\mu}(A) \leq \bar{\mu}(B) + \epsilon$. Since this inequality holds for every positive ϵ, we have $\bar{\mu}(A) \leq \bar{\mu}(B)$.

3. Fix any positive number ϵ. Let $\mathcal{C}_1$ be a covering of A_1 with $L(\mathcal{C}_1) \leq \bar{\mu}(A_1) + \epsilon$; let $\mathcal{C}_2$ be a covering of A_2 with $L(\mathcal{C}_2) \leq \bar{\mu}(A_2) + \epsilon/2$; let $\mathcal{C}_3$ be a covering of A_3 with $L(\mathcal{C}_3) \leq \bar{\mu}(A_3) + \epsilon/4$; etc. Let $\mathcal{C}$ denote the collection of open intervals consisting of all those in $\mathcal{C}_1$, together with all those in $\mathcal{C}_2$, together with all those in $\mathcal{C}_3$, etc. Then $L(\mathcal{C}) = L(\mathcal{C}_1) + L(\mathcal{C}_2) + \ldots$ But $\mathcal{C}$ is a covering of $\cup A_i$. Hence, $\bar{\mu}(\cup A_i) \leq L(\mathcal{C}) = L(\mathcal{C}_1) + L(\mathcal{C}_2) + \ldots \leq \bar{\mu}(A_1) + \epsilon + \bar{\mu}(A_2) + \epsilon/2 + \bar{\mu}(A_3) + \epsilon/4 + \ldots = \Sigma \bar{\mu}(A_i) + 2\epsilon$. Since $\bar{\mu}(\cup A_i) \leq \Sigma \bar{\mu}(A_i) + 2\epsilon$ for every positive ϵ, we must have $\bar{\mu}(\cup A_i) \leq \Sigma \bar{\mu}(A_i)$.

Note that the three properties above for outer measure are similar to, but rather weaker than, the additivity property for a measure space. For example, in the third property, we do not require that the A_i be disjoint, while the conclusion is inequality rather than equality. Note also that there is apparently no hope of even stating the last two properties for a general measure space, for we do not know what "$\leq$" means in a general complete abelian topological group.

We are now ready, finally, to specify our measure space. The idea is that, for any measurable set, its measure will be its outer measure, but that not every (sub)set (of the reals) will be measurable. We choose for the measurable sets those which "when used to split any

ather set into two pieces, does so in such a way that, for those pieces, outer measure is additive". More precisely,

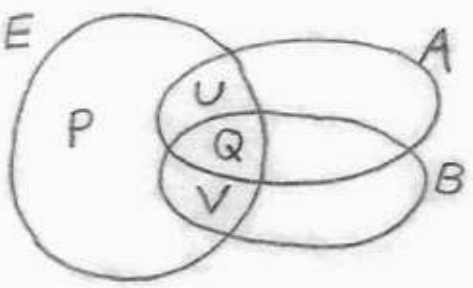
Theorem 7. Let X be the set of reals. Let the measurable subsets of X consist of those subsets A such that $\bar{\mu}(E) = \bar{\mu}(E \cap A) + \bar{\mu}(E \cap A^c)$ for every subset E of X. Let M be the complete abelian topological group of reals. For A measurable, let $\mu(A) = \bar{\mu}(A)$. Then $(X, \mathcal{M}, M, \mu)$ is a measure space.

Proof: We must verify the conditions for a measurable space, and those for a measure space. It is convenient to proceed in steps as follows.

i) The empty set is measurable, and $\mu(\phi) = 0$. For $A = \phi$, $\bar{\mu}(E) = \bar{\mu}(E \cap A) + \bar{\mu}(E \cap A^c)$ becomes $\bar{\mu}(E) = \bar{\mu}(\phi) + \bar{\mu}(E)$. But, by property 1 on the previous page, this holds for every E. Hence, ϕ is measurable. Again by property 1, $\mu(\phi) = \bar{\mu}(\phi) = 0$.

2i) The complement of a measurable set is measurable. Let A be measurable, so $\bar{\mu}(E) = \bar{\mu}(E \cap A) + \bar{\mu}(E \cap A^c)$ for every E. Replacing A by A^c in this equation, we obtain the same equation. Hence, A^c is also measurable.

iii) The union of any two measurable sets is measurable. Let A and B be measurable, and let E be any subset of X. Let P, Q, U, and V denote the four sets indicated in the figure on the right. Since A is measurable, we have $\bar{\mu}(E) = \bar{\mu}(P \cup V) + \bar{\mu}(Q \cup U)$.



Since B is measurable, we have $\bar{\mu}(P \cup V) = \bar{\mu}(P) + \bar{\mu}(V)$. Substituting the second equation in the first, $\bar{\mu}(E) = \bar{\mu}(P) + \bar{\mu}(V) + \bar{\mu}(Q \cup U)$. But, since A is measurable, $\bar{\mu}(V \cup Q \cup U) = \bar{\mu}(V) + \bar{\mu}(Q \cup U)$, and so the last equation becomes $\bar{\mu}(E) = \bar{\mu}(P) + \bar{\mu}(V \cup Q \cup U)$. But $P = E \cap (A \cup B)^c$, while $V \cup Q \cup U = E \cap (A \cup B)$. We conclude that $\bar{\mu}(E) = \bar{\mu}(E \cap (A \cup B)^c) + \bar{\mu}(E \cap (A \cup B))$ for every E. Hence, $A \cup B$ is measurable.

iv) For $A_1, A_2, \ldots$ disjoint and measurable, $\cup A_i$ is also measurable, and $\sum \mu(A_i) = \mu(\cup A_i)$. Let $A_1, A_2, \ldots$ be disjoint and measurable, and let E be any subset of X. Consider the following string of inequalities:

$$\begin{aligned}
\bar{\mu}(E) &= \bar{\mu}((\overset{n}{\underset{1}{\cup}} A_i) \cap E) + \bar{\mu}((\overset{n}{\underset{1}{\cup}} A_i)^c \cap E) \\
&= \overset{n}{\underset{1}{\Sigma}} \bar{\mu}(A_i \cap E) + \bar{\mu}((\overset{n}{\underset{1}{\cup}} A_i)^c \cap E) \\
&\geq \overset{n}{\underset{1}{\Sigma}} \bar{\mu}(A_i \cap E) + \bar{\mu}((\overset{\infty}{\underset{1}{\cup}} A_i)^c \cap E) \\
&\geq \overset{\infty}{\underset{1}{\Sigma}} \bar{\mu}(A_i \cap E) + \bar{\mu}((\overset{\infty}{\underset{1}{\cup}} A_i)^c \cap E) \\
&\geq \bar{\mu}((\overset{\infty}{\underset{1}{\cup}} A_i) \cap E) + \bar{\mu}((\overset{\infty}{\underset{1}{\cup}} A_i)^c_c \cap E) \\
&\geq \bar{\mu}(E).
\end{aligned}$$

We claim that the assertion represented by each line is true. The first assertion follows from step iii) above: the union of any finite number of measurable sets is measurable. The second assertion follows from the fact that each A_i is measurable. The third assertion follows from property 2 on page 20, together with the fact that $(\overset{n}{\underset{1}{\cup}} A_i)^c \cap E$ contains $(\overset{\infty}{\underset{1}{\cup}} A_i)^c \cap E$ as a subset. The fourth assertion follows from the fact that the third holds for all n. The fifth assertion follows from property 3 on page 20. The sixth assertion also follows from property 3, together with the fact that $((\overset{\infty}{\underset{1}{\cup}} A_i) \cap E) \cup ((\overset{\infty}{\underset{1}{\cup}} A_i)^c \cap E) = E$.

Since $\bar{\mu}(E)$ appears at each end of this string of inequalities, each inequality must be an equality. This is true, in particular, for the last inequality. Replacing "$\geq$" there by "=", we obtain precisely the statement that $\cup A_i$ is measurable.

The inequalities (now equalities) above hold for every E. Choose $E = \cup A_i$. Then the fourth line becomes $\overset{\infty}{\underset{1}{\Sigma}} \bar{\mu}(A_i) + \bar{\mu}(\phi)$, while $\bar{\mu}(E)$ becomes $\bar{\mu}(\cup A_i)$. Hence, $\Sigma \bar{\mu}(A_i) = \bar{\mu}(\cup A_i)$, and so, since $\mu = \bar{\mu}$ on measurable sets, this holds also with the bars removed.

v) The union of any countable collection of measurable sets is measurable. Let $A_1, A_2, \ldots$ be measurable, and set $B_1 = A_1$, $B_2 = A_2 \cap A_1^c$, $B_3 = A_3 \cap A_2^c \cap A_1^c$, $B_4 = A_4 \cap A_3^c \cap A_2^c \cap A_1^c$, etc. Then the B's are disjoint, while, by conditions ii) and iii) above, each B_i is measurable. Hence, by step iv) above, $\cup B_i$ is measurable. But $\cup B_i = \cup A_i$.

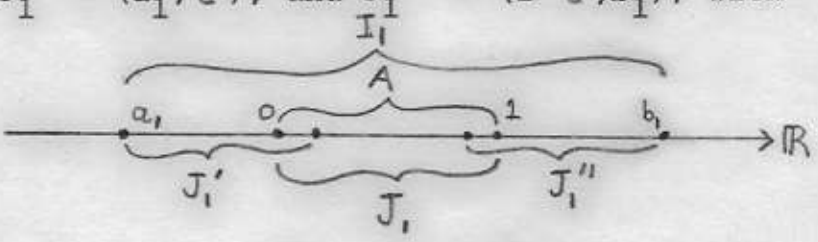
The proof is completed by noting that the three conditions for a measurable space and the two conditions for a measure space are immediate consequences of steps i)-v). /

There are two tricky parts in the proof. The proof of step iii) is basically straightforward: one splits into pieces, and then reassembles. One must, however, take a little care to avoid subtracting outer measures, for $\infty - \infty = 0$. The string of inequalities in step iv) is, for me, rather pretty - but also somewhat mysterious. It might be of interest to try to understand this argument better, so that, e.g., one could remember it.

Thus, our criterion of measurability for A is that $\bar{\mu}(E) = \bar{\mu}(E \cap A) + \bar{\mu}(E \cap A^c)$ for every E. Remarkably enough, we obtain, with this criterion, a measure space. Note that we actually prove more than is needed in the theorem above: we in fact show that, for any partition $A_1, A_2, \ldots$ of A, $\mu(A) = \Sigma \mu(A_i)$. We have seen already that it would not be advisable to incorporate such a property in the definition of a general measure space. Why does it happen to hold in this example? The reason, of course, is that outer measures are non-negative. [For example, Theorem 3 holds in $\mathbb{R}^*$ if we replace the condition that the original series converge by the condition that its terms be non-negative.] The definition of a measure space is what it is, essentially, because we don't know what "non-negative" means in a general complete abelian topological group. It is also worth noting that the only properties of outer measure we used in Theorem 7 are those three listed on page 20.

The measure space obtained in Theorem 7 is called (the) Lebesque measure (space).

Example. Let A be the open interval (0,1). We show that A is measurable. Let E be any subset of the reals, and fix any positive number ϵ. Let $\mathcal{C}$ be a covering $I_1 = (a_1,b_1), I_2 = (a_2,b_2), \ldots$ of E with $L(\mathcal{C}) \leq \bar{\mu}(E) + \epsilon$. With the interval I_1, associate the three intervals $J_1 = I_1 \cap A$, $J_1' = (a_1, \epsilon)$, and $J_1'' = (1-\epsilon, b_1)$; with the interval I_2 the three intervals $J_2 = I_2 \cap A$, $J_2' = (a_2, \epsilon/2)$, $J_2'' = (1-\epsilon/2, b_2)$, etc. [We adopt the convention that an interval for which the first entry exceeds the second is empty.]



Let $\mathcal{D}$ denote the collection $J_1, J_2, \ldots$ of open intervals, and $\mathcal{D}'$ the collection J_1', J_1'', J_2', J_2'', etc.. Then we have

$L(\mathcal{D}) + L(\mathcal{D}') \leq L(\mathcal{C}) + 4\epsilon$. But $\mathcal{D}$ covers $E \cap A$, and $\mathcal{D}'$ covers $E \cap A^c$, whence $\bar{\mu}(E \cap A) \leq L(\mathcal{D})$ and $\bar{\mu}(E \cap A^c) \leq L(\mathcal{D}')$. Hence, $\bar{\mu}(E \cap A) + \bar{\mu}(E \cap A^c) \leq L(\mathcal{D}) + L(\mathcal{D}') \leq L(\mathcal{C}) + 4\epsilon \leq \bar{\mu}(E) + 5\epsilon$. Since this holds for every ϵ, we have $\bar{\mu}(E \cap A) + \bar{\mu}(E \cap A^c) \leq \bar{\mu}(E)$. But the reverse inequality always holds, by property 2 on page 20. Hence, $\bar{\mu}(E \cap A) + \bar{\mu}(E \cap A^c) = \bar{\mu}(E)$. Since E is arbitrary, A is measurable. Of course, $\mu(A) = 1$.

By the same argument as in this example, every open interval in the reals is measurable. Since we have a measurable space, so are complements of open intervals, and countable intersections and unions of open intervals and their complements. [It turns out, although we shall not show it, that there are actually measurable sets which cannot be obtained as a countable union of countable intersections of open intervals and their complements.] In particular, every countable subset of the reals is measurable (and has measure zero). In short, every subset of the reals that one can think of offhand is measurable. One might even imagine that every subset of the reals is measurable. It turns out that this is not true, but it is "nearly true": the construction of a non-measurable subset of the reals is subtle. There follows one such construction.

Example. For any numbers a and b in the open interval (0,1), write $a \approx b$ if $a - b$ is rational. This is an equivalence relation. Let A be a subset of the reals consisting of exactly one element from each equivalence class. (Thus, no two elements of A differ by a rational, while any number in (0,1) differs by a rational from some element of A.) We show that the assumption that this A is measurable (say, with $\mu(A) = \alpha$, an element of R*) leads to a contradiction. For each rational r in [0,1), let A_r consist of all numbers a in (0,1) with (a+r) (mod 1) in A. Since, for each r, A_r is the union of the disjoint sets $(A+r) \cap (0,1)$ and $(A+r-1) \cap (0,1)$, and since Lebesque measure is translation invariant (i.e., $\mu(B + r) = \mu(B)$ for any measurable B), each A_r must be measurable, with $\mu(A_r) = \alpha$. But the A_r's are disjoint (since no two elements of A differ by a rational), and their union is (0,1) (since every element of (0,1) differs by a rational from some element of A). Hence, the A_r's form a partition of (0,1), whence we must have $1 = \mu(0,1) = \Sigma \mu(A_r) = \alpha + \alpha + \alpha \ldots$ But this is a contradiction, for, for $\alpha = 0$, the right side is zero, while, for $\alpha > 0$, the right side is ∞. In particular, outer measure alone does not yield directly a measure space.

7. Null Sets; Completeness

We now consider sets which, as far as measure is concerned, "don't count".

Let $X, \mathcal{M}, M, \mu$ be a measure space. A measurable subset A of X is said to be null if $\mu(B) = 0$ for every measurable subset B of A. In particular, every null set has measure zero.

Example. In Lebesque measure, measurable A is null if and only if $\mu(A) = 0$. Proof: Clearly, if A is null, then $\mu(A) = 0$. Suppose, conversely, that $\mu(A) = 0$. Then, for any measurable subset B of A, we have $\mu(B) + \mu(B^c \cap A) = \mu(A) = 0$. Since Lebesque measure is non-negative, this implies $\mu(B) = 0$. We conclude that A is null.
Example. In the example on page 15, A is null if and only if $m_i = 0$ for each i in A.

The notion of a "null set" in a general measure space replaces that of "set of measure zero" in Lebesque measure, a replacement suggested by the first example. The second example shows that "null" is not in general the same as "measure zero".

Several properties of null sets are immediate from the definition. Clearly, any measurable subset of a null set is null. Furthermore, for A measurable and B null, $\mu(A \cup B) = \mu(A)$. [Proof: $\mu(A \cup B) = \mu(A) + \mu(B \cap A^c)$. But $B \cap A^c$, as a subset of null B, has measure zero.] Similarly: for A measurable and B null, $\mu(A \cap B^c) = \mu(A)$. The following property is only slightly more subtle: the union of a countable collection of null sets is null. [Proof: Let $A_1, A_2, \ldots$ be null. We may take them to be disjoint, for $A_1, A_2 \cap A_1^c, A_3 \cap A_2^c \cap A_1^c, \ldots$ are disjoint null sets whose union is the same as the union of the A_i. Set $A = \cup A_i$. Choosing a fine refinement of the partition $A_1, A_2, \ldots$ of A, we see that $\mu(A) = 0$. Let B be a measurable subset of A, so $\mu(B)$ is finite. Then $\mu(B) = \mu(B \cap A_1) + \mu(B \cap A_2) + \ldots = 0 + 0 + \ldots = 0$. Hence, A is null.]

It is certainly possible to have the situation in which A is a null set, B is some subset of A, but B is not even measurable.

Example. Let X be any set with at least two elements. Let the measurable subsets of X be X and $\emptyset$. Let $\mu(X) = \mu(\emptyset) = 0$. Then X is null, but any subset of X neither X nor $\emptyset$ is non-measurable.

For B a subset of null A, "B might just as well be measurable, for we know already what its measure should be, namely zero". This remark suggests the following definition: a measure space is said to be complete if every subset of a null set is measurable.

Example. The example on page 15 is complete, for there every set is measurable. The last example on the previous page is, of course, not complete.

Example. Lebesque measure is complete. Proof: Let A be a set (in Lebesque measure) of measure zero, and let B be any subset of A. We show that B is measurable. Let E be any set. Then, by property 3 on page 20, $\bar{\mu}(E) \leq \bar{\mu}(E \cap B) + \bar{\mu}(E \cap B^c)$. By property 2 on page 20, $\bar{\mu}(E) \geq \bar{\mu}(E \cap B^c)$. Again by property 2 (together with the facts that $E \cap B$ is a subset of A, and $\bar{\mu}(A) = 0$), $\bar{\mu}(E \cap B) = 0$. But these three equations imply $\bar{\mu}(E) = \bar{\mu}(E \cap B) + \bar{\mu}(E \cap B^c)$. Since this holds for every E, B is measurable.

We regard a measure space which is not complete as one which "could have been endowed with more measurable sets from the beginning". The next theorem not only amplifies this viewpoint, but also shows how the restoration of the missing measurable sets is carried out.

Theorem 8. Let $X, \mathcal{M}, M, \mu$ be a measure space. Let $\hat{\mathcal{M}}$ consist of subsets K of X for which there exist A and B in $\mathcal{M}$ with $A \subset K \subset B$ and with $B \cap A^c$ null. For such K, set $\hat{\mu}(K) = \mu(A)$. Then $\hat{\mu}$ is well-defined, and $X, \hat{\mathcal{M}}, M, \hat{\mu}$ is a complete measure space.

Proof: We first show that $\hat{\mu}$ is well-defined, i.e.: if $A \subset K \subset B$ and $A' \subset K \subset B'$, with A, A', B, B' in $\mathcal{M}$, and with $B \cap A^c$ and $B' \cap A'^c$ null, then $\mu(A) = \mu(A')$. But this is immediate from the fact that $A = A' \cup Q$, where $P = A' \cap A^c \subset B \cap A^c$ and $Q = A \cap A'^c \subset B' \cap A'^c$ are null.

We next show that $X, \hat{\mathcal{M}}$ is a measurable space. Since $\emptyset \subset \emptyset \subset \emptyset$, $\emptyset$ is in $\hat{\mathcal{M}}$. Let K be in $\hat{\mathcal{M}}$, so $A \subset K \subset B$ with $B \cap A^c$ null. Then $B^c \subset K^c \subset A^c$ with $A^c \cap B^{cc} = B \cap A^c$ null, whence K^c is in $\hat{\mathcal{M}}$. Let $K_1, K_2, \ldots$ be in $\hat{\mathcal{M}}$, say, with $A_1 \subset K_1 \subset B_1$, $A_2 \subset K_2 \subset B_2, \ldots$, and $B_1 \cap A_1^c, \ldots$ null. Then $\cup A_i \subset \cup K_i \subset \cup B_i$, with $(\cup B_i) \cap (\cup A_i)^c \in \cup (B_i \cap A_i^c)$ null. Hence, $\cup K_i$ is in $\hat{\mathcal{M}}$.

We next show that $X, \hat{\mathcal{M}}, M, \hat{\mu}$ is a measure space. Clearly, $\hat{\mu}(\emptyset) = 0$. Note also that any set in $\mathcal{M}$ is also in $\hat{\mathcal{M}}$. Now let K be in $\hat{\mathcal{M}}$, say with $A \subset K \subset B$ and $B \cap A^c$ null. Let $A_1, A_2, \ldots$ be any

μ-fine partition of A. Then $K \cap A^c$, A_1, A_2, . . . is a partition of K in $\hat{\mathcal{M}}$. We show that it is $\hat{\mu}$-fine. Let K_1, K_2, . . . be a $\hat{\mathcal{M}}$-refinement, and let $C_i \subset K_i \subset D_i$, with each $E_i \cap C_i^c$ null. Then $\mu(C_i) = 0$ for each i with $K_i \subset K \cap A^c$. Furthermore, the $\mathcal{M}$-partition of A consisting of all C_i with i such that $K_i \subset A$, together with all sets of the form $A_i \cap (\cup_j C_j)^c$ (i = 1,2,...) is a $\mathcal{M}$-refinement of the partition A_1, A_2, . . . Hence, $\Sigma \mu(C_i) = \mu(A)$, whence $\Sigma \hat{\mu}(K_i) = \hat{\mu}(K)$.

Finally, we show that the measure space $X, \hat{\mathcal{M}}, M, \hat{\mu}$ is complete. Let K (with $A \subset K \subset B$, and $B \cap A^c$ null) be $\hat{\mu}$-null, and let L be a subset of K. Since every μ-measurable subset C of A is a $\hat{\mu}$-measurable subset of K, and since K is $\hat{\mu}$-null, $\mu(C) = 0$ for every such C. That is, A is μ-null. But $B = A \cup (B \cap A^c)$, and so B, as the union of two μ-null sets, is μ-null. But $\emptyset \subset L \subset B$, while, as we have just shown, $B \cap \emptyset^c = B$ is μ-null. Hence, L is in $\hat{\mathcal{M}}$. We have shown that every subset of a $\hat{\mu}$-null set is $\hat{\mu}$-measurable, whence $X, \hat{\mathcal{M}}, M, \hat{\mu}$ is complete. /

Thus, starting with any measure space $X, \mathcal{M}, M, \mu$, one obtains, by the construction of Theorem 8, a complete measure space. This $X, \hat{\mathcal{M}}, M, \hat{\mu}$ is called the completion of $X, \mathcal{M}, M, \mu$. Of course, any measurable set in the original measure space is also measurable in its completion: in this sense, the completion merely "calls a few more sets measurable". Note also (immediately from the statement of Theorem 8) that the completion of a complete measure space is that same measure space. That is, one gets nothing new by taking the completion over and over.

Example. The completion of the measure space in the last example on page 25 is this: every subset of X is measurable, and has measure zero.

Example. Let X be the set of reals, let the measurable sets be $\emptyset$ and X, and let $\mu(\emptyset) = 0$ and $\mu(X) = \infty$. This measure space is complete.

This last example shows that, even with a complete measure space, it may still be possible to "call more sets measurable, and assign them measures in a consistent way", for one could, in that example, "expand the measure space to Lebesque measure".

8. Small Sets

We introduce in this section the comparative version of the superlative of the previous section. That is to say, small sets will "contain very little in the way of measure", while null sets "contain nothing in the way of measure".

Let $X, \mathcal{M}, M, \mu$ be a measure space. A measurable set A will be called $\mathcal{U}$-small if, for every measurable subset B of A, $\mu(B)$ is in U. Thus, in particular, every U-small set has finite measure. It is immediate from the first condition on page 7 that a set is null if and only if it is U-small for every U. Of course, every measurable subset of a U-small set is U-small. We claim, furthermore, that, if A is U-small and B is V-small, then $A \cup B$ is (U + V)-small. Proof: For C a measurable subset of $A \cup B$, we have $\mu(C) = \mu(C \cap A) + \mu(C \cap B \cap A^c)$. But $C \cap A$ is a subset of A, and $C \cap B \cap A^c$ is a subset of B.]

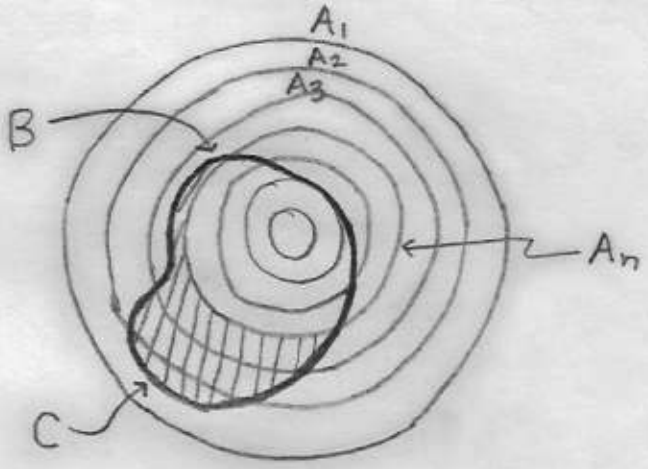
Example. In Lebesque measure, let A be a measurable set with $\alpha = \mu(A) \neq 0$ finite. Let U be the 0-neighborhood $(-\epsilon, \alpha)$, where ϵ is any positive number. Then A is U-small.

The basic result about small sets states that "if a collection of sets contains small sets in the set-theoretic sense, then it also contains small sets in the measure-theoretic sense".

Theorem 9. Let $X, \mathcal{M}, M, \mu$ be a measure space, and let U be a 0-neighborhood. Let $A_1, A_2, \ldots$ be a countable collection of measurable sets such that i) the intersection of any two of the A's is another, ii) the intersection of all the A's is empty, and iii) at least one A_i has finite measure. Then at least one A_i is U-small.

Proof: We may assume that $A_1 \supset A_2 \supset A_3 \ldots$, for each of A_1, $A_1 \cap A_2$, $A_1 \cap A_2 \cap A_3, \ldots$ contains its successor as a subset, while, by condition i) above, each set in this sequence is one of the A_i's. We may also assume that each A_i has finite measure, for, for $A_1 \supset A_2 \supset \ldots$, condition iii) above implies that all the A_i beyond a certain one have finite measure. Make these assumptions.

We suppose that no A_i is U-small, and obtain a contradiction. Let V be a 0-neighborhood with V + V a subset of U. Since A_1 is not U-small, there is a measurable subset B of A_1, with $\mu(B)$ not in U. Since B has finite measure, $\mu(B) = \mu(B \cap A_1 \cap A_2^c) + \mu(B \cap A_2 \cap A_3^c) + \mu(B \cap A_3 \cap A_4^c) + \ldots$, where we have used condition iii). Since



this series converges, there is some integer n such that $\sum_{i>n} \mu(B \cap A_i \cap A_{i+1}^c)$ is in V. Set $C = B \cap A_1 \cap A_n^c$. Then, since $\mu(B) = \mu(C) + \sum_{i>n} \mu(B \cap A_i \cap A_{i+1}^c)$, since the second term in is V, since V+V is a subset of U, and since $\mu(B)$ is not in U, $\mu(C)$ cannot be in V.

Since A_n is not U-small, there is a measurable subset B' of A_n with $\mu(B')$ not in U. Repeating the argument above, we obtain an integer $n' > n$, and a measurable subset C of $A_n \cap A_{n'}^c$ with $\mu(C)$ not in V. Continue in this way, to obtain a collection C, C', C'', . . . of disjoint measurable subsets of A_1, each of which has measure not in V. But $\mu(C \cup C' \cup C'' \cup \ldots) = \mu(C) + \mu(C') + \mu(C'') + \ldots$, a contradiction, since the left side (as the measure of a measurable subset of A_1, with $\mu(A_1)$ finite) is finite, while the right side (since none of $\mu(C)$, $\mu(C')$, . . . is in V) is infinite. This contradiction establishes that some A_i is U-small. /

It is natural to ask whether there might be some stronger version of Theorem 9. The theorem is clearly false if either of conditions i) or ii) are eliminated. It also turns out to be false without condition iii):

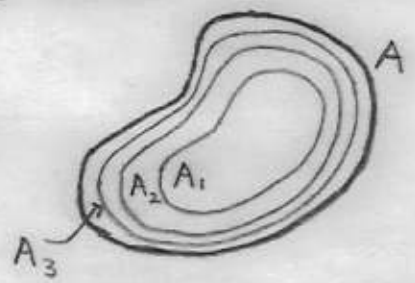
Example. Let X be the set of positive integers, let all subsets of X be measurable, and let $\mu(A)$ be zero if A is empty, and ∞ otherwise. For each i, let A_i consist of all integers i or greater. Then all the hypotheses of Theorem 9 except condition iii) are satisfied. But $\mu(A_i) = \infty$ for each i, and so no A_i is U-small.

A more innocuous-looking condition is our requirement in Theorem 9 that there be only a countable collection of A's. This, too, is necessary:

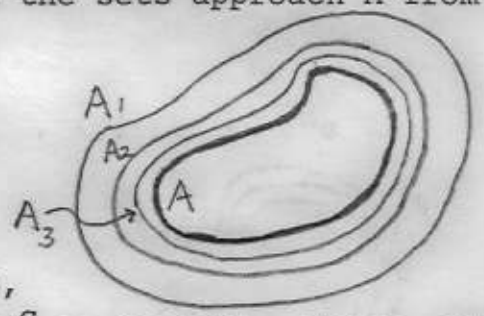
Example. In Lebesque measure, consider the collection of all subsets A of (0,1) with $(0,1) \cap A^c$ finite. The (uncountable) collection of such A's satisfies all the other conditions of Theorem 9. But, for every such A, $\mu(A) = 1$, and so no such A is (-1/2,1/2)-small.

One uses Theorem 9 whenever one wants to show that "nearby sets have nearby measures". We give two very similar examples, both of which may be regarded as amplifications of the motivating remarks on page 5.

Let $A_1 \subset A_2 \subset \ldots$, let $A = \cup A_i$, and let A have finite measure. We show that the sequence $\mu(A_1)$, $\mu(A_2)$, . . . in M converges to $\mu(A)$. The sets $A \cap A_1^c$, $A \cap A_2^c$, . . . satisfy the conditions of Theorem 9. Hence, by that theorem, given any 0-neighborhood U, there is an n with $A \cap A_n^c$ U-small. Let $i \geq n$. Then $\mu(A) = \mu(A_i) + \mu(A \cap A_i^c)$, while, since $A \cap A_i^c$ is a subset of U-small $A \cap A_n^c$, $\mu(A \cap A_i^c)$ is in U. That is to say, $\mu(A_i)$ is in $\mu(A) - U$. We have shown that, for every 0-neighborhood U, there is an n such that $\mu(A_i)$ is in $\mu(A) - U$ whenever $i \geq n$. But this is just the statement that the sequence $\mu(A_1)$, $\mu(A_2)$, . . . converges to $\mu(A)$.



One can do the same thing when the sets approach A from the other direction. Let $A_1 \supset A_2 \supset \ldots$ each have finite measure, and set $A = \cap A_i$. Then $\mu(A_1)$, $\mu(A_2)$, . . . converges to $\mu(A)$. Proof: The sets $A_1 \cap A^c$, $A_2 \cap A^c$, . . . satisfy the conditions of Theorem 9, whence, for any 0-neighborhood U, some $A_n \cap A^c$ is U-small. Hence, as above, $\mu(A_i)$ is in $\mu(A) + U$ whenever $i \geq n$.



We shall many times find it necessary to use these two convergence properties of measures.

9. Sets of Finite Measure

The previous two sections, together with this one, form a sequence in which we study sets with increasingly rich internal structure. A complete table, including also sets with infinite measure, might be the following:

Types of Measurable Sets

Name	Defining Property
Null	Every measurable subset has measure zero.
Finite	Has finite measure.
Sigma-Finite	Is the union of a countable collection of measurable sets of finite measure.
Essentially Infinite	Is not the union of any countable collection of measurable sets with finite measure.

The last two entries in the table introduce new definitions. It is easily seen from our previous examples that all four types of measurable sets can occur. Let A be a sigma-finite set having infinite measure (e.g., A = X in Lebesque measure). Then, since there is a countable collection of measurable sets each of which has finite measure, and whose union is A, there is a fine partition $A_1, A_2, \ldots$ of A with each $\mu(A_i)$ finite. For this partition, we have $\mu(A) = \sum \mu(A_i) = \infty$. Thus, we can regard this set A as having infinite measure because "it was forced to do so by what the measures of the finite sets are". The essentially infinite sets, on the other hand, have infinite measure "because they felt like it". Note that the following general statements about the types are true: i) every measurable set is of some type, ii) a measurable set of some type has the property that every measurable subset is of that type or an earlier one, iii) the union of two measurable sets of some type is of that type. Note also that the union of a countable collection of null sets is null; of a countable collection of finite sets is sigma-finite; and of a countable collection of sigma-finite sets is sigma-finite. We shall call a measure space null, finite, sigma-finite, or essentially infinite according to the type of the measurable set X.

In this section, we consider the third and fourth types in the table above - specifically, their relation to the second type.

It is convenient to begin with a definition.

A pre-measure space consists of four things - i) a measurable space $X, \mathcal{M}$, ii) a collection $\mathcal{F}$ of measurable sets, iii) a complete abelian topological group M, and iv) a mapping ν from the set $\mathcal{F}$ to the set M - subject to the following four conditions:

1. ϕ is in $\mathcal{F}$, and $\nu(\phi) = 0$.
2. If A is in $\mathcal{F}$, and B is a measurable subset of A, then B is in $\mathcal{F}$.
3. If A and B are in $\mathcal{F}$, then so is $A \cup B$.
4. For any A in $\mathcal{F}$, and any partition $A_1, A_2, \ldots$ of A (whence, by 2, each A_i is in $\mathcal{F}$), $\nu(A) = \Sigma \nu(A_i)$.

This definition is, of course, to be compared with that of a measure space. First, a pre-measure space requires not only a measurable space $X, \mathcal{M}$, but also a certain collection $\mathcal{F}$ of measurable sets. The first two conditions say essentially that "sufficiently small" (in the set-theoretic sense) measurable sets are included in $\mathcal{F}$. The third condition requires that one can "build up larger sets in $\mathcal{F}$ from those already in $\mathcal{F}$". The crucial point here is that we do not require that countable unions of sets in $\mathcal{F}$ be in $\mathcal{F}$. The second difference is that ν is only defined on those measurable sets which happen to be in $\mathcal{F}$, and not on all measurable sets. Third, ν takes only finite values, i.e., it maps to M rather than M*. The fourth condition is now analogous to the second condition for a measure space: we don't have to worry about finer partitions since $\nu(A)$ is always finite.

The following example will be seen as the reason why this definition is what it is.

Example. Let $X, \mathcal{M}, M, \mu$ be any measure space. Denote by $\mathcal{F}$ the collection of all measurable sets A with $\mu(A)$ finite. For any such A, set $\nu(A) = \mu(A)$. We claim that $X, \mathcal{M}, \mathcal{F}, M, \nu$ is a pre-measure space. Indeed, the first condition is obvious, the second is immediate from Theorem 4, the third is immediate from the fact that the union of two sets with finite measure has finite measure, and the fourth is immediate from Theorem 6.

One might imagine that every pre-measure space arises as in the previous example. Had this been the case, we would have suppressed the definition of a measure space, replacing it with that of a pre-measure space. Unfortunately,

Example. Let $X, \mathcal{M}, M, \mu$ be Lebesque measure. Denote by $\mathcal{F}$ the collection of all measurable, bounded subsets of X (recalling that a subset of the reals is said to be bounded if there is some number N such that $|r| \leq N$ for every r in that subset). For any such A, set $\nu(A) = \mu(A)$ (noting that $\mu(A)$ is finite for every bounded, measurable A). Then $X, \mathcal{M}, \mathcal{F}, M, \nu$ is a pre-measure space, as is immediate from the previous example, together with the observations that the empty set is bounded, that every subset of a bounded set is bounded, and that the union of two bounded sets is bounded. We claim, furthermore, that there is no measure space from which this example arises by the construction of the previous example, an assertion of which one easily convinces himself.

This last example is clearly contrived: we simply left some sets of finite measure out of $\mathcal{F}$, but not so many as to violate the conditions for a pre-measure space. We make this observation more precise as follows.

Theorem 10. Let $X, \mathcal{M}, \mathcal{F}, M, \nu$ be a pre-measure space. For any A in $\mathcal{M}$,

i) If A is in $\mathcal{F}$, set $\mu(A) = \nu(A)$.

ii) If A is not in $\mathcal{F}$, if there is a partition $A_1, A_2, \ldots$ of A with each A_i in $\mathcal{F}$, and if for every such partition $\Sigma \nu(A_i)$ is finite, set $\mu(A) = \Sigma \nu(A_i)$.

iii) If A is not in $\mathcal{F}$, if there is a partition $A_1, A_2, \ldots$ of A with each A_i in $\mathcal{F}$ and with $\Sigma \nu(A_i)$ infinite, set $\mu(A) = \infty$.

iv) If A is not in $\mathcal{F}$, and if there is no partition $A_1, A_2, \ldots$ of A with each A_i in $\mathcal{F}$, set $\mu(A) = \infty$.

Then μ is well-defined, and $X, \mathcal{M}, M, \mu$ is a measure space.

Proof: We first show that μ is well-defined. Since every A is in precisely one of the four classes above, we have only to check case ii). That is, we must show: if $A_1, A_2, \ldots$ and $A_1', A_2', \ldots$ are two partitions of A with each A_i in $\mathcal{F}$, then $\Sigma \nu(A_i) = \Sigma \nu(A_i')$ provided every such sum is finite. Since any two partitions have a

common refinement, we may assume, without loss of generality, that A_1', A_2', . . . is a refinement of A_1, A_2, . . . By condition 4 for a pre-measure space, $\sum \nu(A_i') = \nu(A_j)$ for each j, where the sum on the left is over all i with A_i' a subset of A_j. But now $\sum \nu(A_i') = \sum \nu(A_i)$ is an immediate consequence of Theorem 3.

Finally, we must check the two conditions for a measure space. The first is clear. For the second condition, we must consider the cases when A is in each of the four classes in the statement of the Theorem. Class i) is immediate from condition 4 for a pre-measure space, and class ii) from what we just checked above. Let A be in class iii), and let A_1, A_2, . . . be a partition of A with each A_i in $\mathcal{F}$, and with $\sum \nu(A_i)$ finite. But this partition is fine, for, for A_1', A_2', . . . any refinement, $\sum \nu(A_i') = \infty$ by Theorem 3. For A in class iv), every partition of A is fine, for, if A_1, A_2, . . . is any partition of A, at least one A_i must be in class iv), whence $\sum \nu(A_i) = \infty = \mu(A)$. Hence, $X, \mathcal{M}, M, \mu$ is a measure space. /

Thus, according to Theorem 10, every pre-measure space leads to a measure space. Clearly, if we start with a measure space, take the pre-measure space of sets of finite measure, and then find the corresponding measure space by Theorem 10, we end up with just the original measure space. Note that classes i) and ii) above give the sets of finite measure (the distinction between the two being whether or not that set happened to be included in $\mathcal{F}$), class iii) gives the sigma-finite sets, and class iv) gives the essentially infinite sets. Our two previous examples show that different pre-measure spaces can give rise to the same measure space.

We conclude that "if one knows the measures of enough sets of finite measure, then one knows the whole measure space". A typical application of Theorem 10 (of which we shall see several later) is along the following lines: one has some construction from which one wants to obtain a measure space. Things work nicely for the sets of finite measure, but badly for those of infinite measure. One therefore constructs instead a pre-measure space, and obtains his final measure space using Theorem 10.

10. Measure Spaces From Measure Spaces

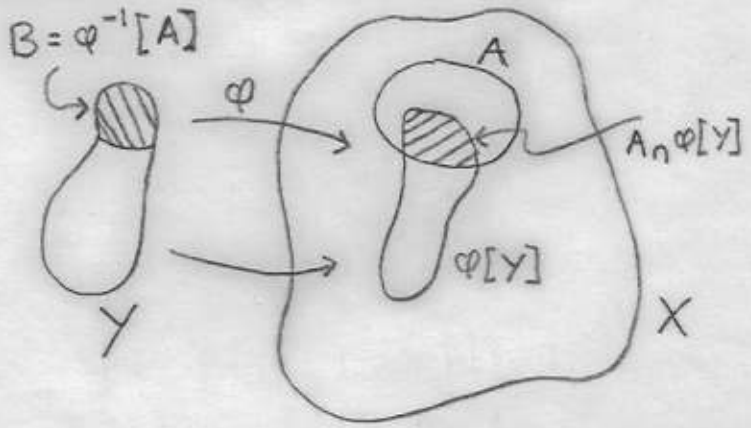
We now begin a program which will occupy our attention for a while: given one or more measure spaces, what additional measure spaces can be constructed therefrom? This question is of interest for at least three reasons. First, the constructions themselves are often used directly in applications. There are essentially two ways in which measure spaces arise in practice: they may be constructed from scratch (e.g., as was Lebesque measure), or they may be constructed from various other measure spaces (such as Lebesque measure) which happen to be available already in one's particular problem. Both methods are common. By organizing, once and for all, the various possibilities for constructing measure spaces from measure spaces, one makes available to himself a large repertoire of constructions for such applications. The second reason is that exposure to these constructions provides one with experience in working with measure spaces, and with a stronger intuition into their structure. We have remarked already that the definition of a measure space is a central one: like all such definitions, it is virtually useless until one has learned to feel comfortable with it. The third reason is that the various constructions we now introduce will provide us with a large fund of illustrative examples of measure spaces, examples on which we can later draw.

11. Changing the Underlying Set

Let $X, \mathcal{M}, M, \mu$ be a measure space. We now consider possibilities for constructing from this measure space another, but with the set X replaced by some other set Y. Of course, one can do nothing with so general a set-up, i.e., without any specified relationship between X and Y. The natural such relationship, and the one we will consider in this section, is the presence of a given mapping from one set to the other.

Let $X, \mathcal{M}, M, \mu$ be a measure space, Y any set, and let φ be a mapping from the set Y to the set X. We wish to base a measure space on Y, and so we must say what the measurable subsets of Y are to be, and what their measures are to be. Denote by $\mathcal{N}$ the collection of all subsets of Y of the form $\varphi^{-1}[A]$, where A is a measurable subset of X. [Recall that $\varphi^{-1}[A]$ is the set of all elements y of Y with $\varphi(y)$ in the set A.] We claim that $Y, \mathcal{N}$ is a measurable space. [Proof: Since $\emptyset = \varphi^{-1}[\emptyset]$, $\emptyset$ is in $\mathcal{N}$. For B in $\mathcal{N}$, say $B = \varphi^{-1}[A]$ with A measurable in X, B^c is also in $\mathcal{N}$, for $B^c = \varphi^{-1}[A^c]$. For $B_1, B_2, \ldots$ $(= \varphi^{-1}[A_1], \varphi^{-1}[A_2], \ldots)$ in $\mathcal{N}$, $\cup B_i = \varphi^{-1}[\cup A_i]$ is in $\mathcal{N}$.]

We next wish to make this $Y, \mathcal{N}$ into a measure space. The following definition might seem reasonable: for $B = \varphi^{-1}[A]$ in $\mathcal{N}$, set $\nu(B) = \mu(A)$. Unfortunately, this is not well-defined, for one could have $B = \varphi^{-1}[A]$ and $B = \varphi^{-1}[A']$ with A and A' different sets (and, in fact, sets with different μ-measures). [For example, let Y have the single element y, and let the measure space be Lebesque measure. Let $\varphi(y) = 0$. Then $Y = \varphi^{-1}[(-1,1)] = \varphi^{-1}[(-2,2)]$.] The problem, of course, is that $\varphi[Y]$ (the set of all points of X of the form $\varphi(y)$ with y in Y) may not be all of X. But $\varphi^{-1}[A]$ depends only on $A \cap \varphi[Y]$, i.e., only on the part of A in $\varphi[Y]$, and not on what elements A may have outside of $\varphi[Y]$. One might therefore imagine modifying our definition of ν as follows: for $B = \varphi^{-1}[A]$, we set $\nu(B) = \mu(A \cap \varphi[Y])$. That is to say, the ν-measure of a subset B of Y is defined as



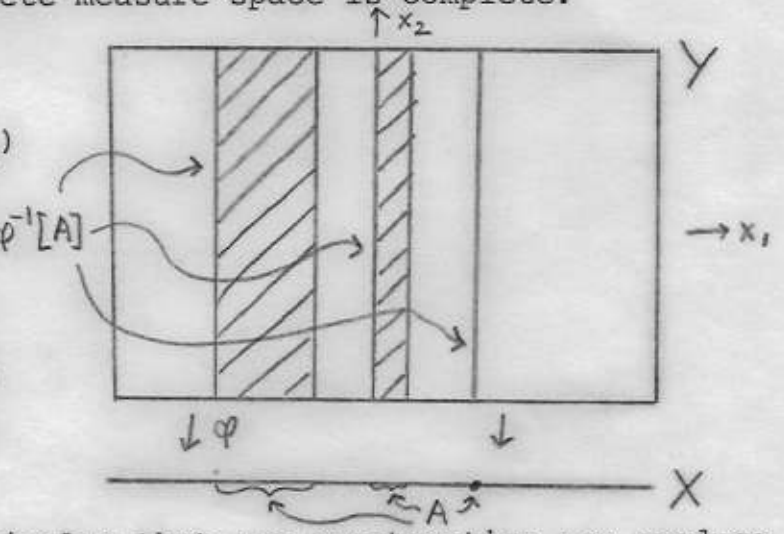
the μ-measure of the part of A in $\varphi[Y]$. Unfortunately, this, too, will not work in general, for we have no guarantee that $A \cap \varphi[Y]$ will be a measurable subset of X (so that we can know what μ of it means). [One could imagine here trying to modify $\mathcal{N}$, to admit only B's with $A \cap \varphi[Y]$ measurable, but this does not lead in general to a measurable space based on Y.]

To make progress, one must assume more. As the discussion above suggests, we should assume that φ is such that $\varphi[Y]$ is measurable in X. Then every B in $\mathcal{N}$ can be written in one and only one way as $B = \varphi^{-1}[A]$ with A a measurable subset of $\varphi[Y]$. Setting $\nu(B) = \mu(A)$, we obtain a well-defined mapping ν from $\mathcal{N}$ to M*. We now claim that $Y, \mathcal{N}, M, \nu$ is a measure space. Indeed, this is immediate from the observation that, for $B = \varphi^{-1}[A]$, where A is a measurable subset of $\varphi[Y]$, $\varphi^{-1}[A_1]$, $\varphi^{-1}[A_2]$, . . . is a fine partition of B whenever A_1, A_2, . . . is a fine partition of A.

A special case arises when the mapping φ is one-to-one. Then, since φ is a one-to-one mapping from Y onto a measurable subset of X, we might as well dispense with φ and regard Y itself as a measurable subset of X. Then $\mathcal{N}$ is the collection of all measurable subsets of Y, and, for such a subset, $\nu(A) = \mu(A)$. Thus, every measurable subset of a measure space gives rise immediately to a new measure space (based on that subset). One says that the second is a <u>measure subspace</u> of the first.

<u>Example</u>. Any measure space obtained, by the construction above, from a sigma-finite (resp., finite) measure space is sigma-finite (resp. finite). Any measure subspace of a complete measure space is complete.

<u>Example</u>. Let $X, \mathcal{M}, M, \mu$ be Lebesque measure, let Y be $\mathbb{R}^2$, and let φ have action $\varphi(x_1, x_2) = x_1$. Then the sets in $\mathcal{N}$ are "strips in the x_2-direction", strips whose cross-section is measurable in the x_1-variable. The ν-measure of such a "strip" is the Lebesque measure of its cross section.



This last example shows in particular that our construction can produce an incomplete measure space from a complete one.

Let $X, \mathcal{M}, M, \mu$ be a measure space, Y and set, and let φ be a mapping from set X to set Y. That is, we have the reverse of the previous situation, and again we want to obtain a measure space based on Y. Denote by $\mathcal{N}$ the collection of all subsets A of Y with $\varphi^{-1}[A]$ measurable in X. Then, by essentially the same remarks as earlier, this $Y, \mathcal{N}$ is a measurable space. To obtain a measure space, we must define the measures of these sets. Consider $\nu(A) = \mu(\varphi^{-1}[A])$. This is certainly well-defined (for the instructions are: if you want to know the ν-measure of A, take φ^{-1} of A, and apply μ to it). Unfortunately, we do not thereby obtain a measure space in general:

Example. Let X be the set of positive integers, all subsets of X measurable, M the reals. Set $\mu(1) = +1$, $\mu(2) = -1$, $\mu(3) = +1$, etc., and obtain a measure as in the example on page 15. Let Y be the set of ~~positive~~ integers, and set $\varphi(1) = \varphi(2) = 1$, $\varphi(3) = \varphi(4) = 2$, etc. Then every subset of Y is in $\mathcal{N}$. For A any finite subset of Y, $\nu(A) = 0$ (for $\varphi^{-1}[A]$ includes as many "+1"'s as "-1"'s). However, $\nu(Y) = \infty$, for $\varphi^{-1}[Y] = X$, and $\mu(X) = \infty$. Hence, $Y, \mathcal{N}, M, \nu$ is not a measure space.

A glance at this last example shows that its mechanism depends essentially on having sets of infinite measure. This observation, in turn, suggests that we must proceed in a more indirect way, using Theorem 10.

Let $X, \mathcal{M}, M, \mu$ be a measure space, Y a set, and φ a mapping from X to Y. Let $\mathcal{N}$ be the collection of subsets A of Y with $\varphi^{-1}[A]$ measurable, so $Y, \mathcal{N}$ is a measurable space. Now let $\mathcal{P}$ denote the collection of all A in $\mathcal{N}$ such that $\varphi^{-1}[A]$ in X has finite μ-measure. For A in $\mathcal{P}$, set $\lambda(A) = \mu(\varphi^{-1}[A])$, so $\lambda(A)$ is always finite. We claim that $Y, \mathcal{N}, \mathcal{P}, M, \lambda$ is a pre-measure space. Proof: 1. Since $\varphi^{-1}[\emptyset] = \emptyset$ is measurable in X, and $\mu(\emptyset) = 0$, $\emptyset$ is in $\mathcal{P}$, and $\lambda(\emptyset) = 0$. 2. Let A be in $\mathcal{P}$, so $\varphi^{-1}[A]$ in X has finite measure. Then, for B a subset of A with $\varphi^{-1}[B]$ measurable in X, $\varphi^{-1}[B]$ has finite measure in X, since $\varphi^{-1}[B]$ is a subset of $\varphi^{-1}[A]$. Hence, B is also in $\mathcal{P}$. 3. For A and B in $\mathcal{P}$, $\varphi^{-1}[A \cup B] = \varphi^{-1}[A] \cup \varphi^{-1}[B]$ has finite measure, whence $A \cup B$ is in $\mathcal{P}$. 4. Let A be in $\mathcal{P}$, and let $A_1, A_2, \ldots$ be a partition. Then $\varphi^{-1}[A_1], \varphi^{-1}[A_2], \ldots$ is a partition of $\varphi^{-1}[A]$, whence $\mu(\varphi^{-1}[A]) = \Sigma \mu(\varphi^{-1}[A_i])$ by Theorem 4. That is to say, $\lambda(A) = \Sigma \lambda(A_i)$.

Thus, we obtain a pre-measure space. We now invoke Theorem 10 to obtain a measure space. Note that this construction gives reasonable answers. (For example, if Y = X, and φ is the identity mapping, the construction yields again the original measure space.) Note also how the construction circumvents the problem illustrated in the example on the previous page:

Example. Let X, $\mathcal{M}$, M, μ, and Y and φ, be as in the example on the previous page. Then $\mathcal{P}$ consists of all finite subsets of Y (for, for any infinite subset A, $\mu(\varphi^{-1}[A])$ is infinite). For any finite A, $\lambda(A) = 0$. This is ourepre-measure space. We now apply Theorem 10. If A is any finite subset of Y, then it is in class i) of that Theorem, whence $\nu(A) = 0$. If A is any infinite subset of Y, then it is in class ii) of that theorem, whence $\nu(A) = 0$. Thus, in the final measure space Y, $\mathcal{N}$, M, ν, every set has measure zero.

To summarize, mappings from Y to X yield a measure space when $\varphi[Y]$ is measurable in X; mappings from X to Y always yield a measure space, but one must proceed via Theorem 10.

12. Homomorphisms

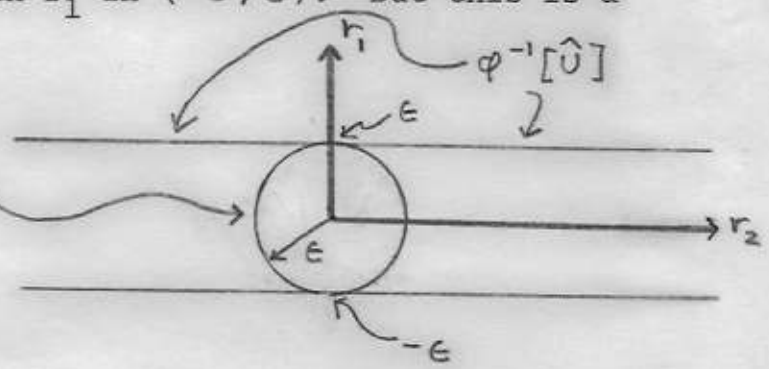
In the previous section, we considered the possibilities for changing the underlying set of a measure space. We next wish to consider the possibilities for changing the group. As a prerequisite for this study, we require the notion of a "structure-preserving mapping" from one complete abelian topological group to another, a notion we introduce in the present section.

Let M and $\hat{M}$ be complete abelian topological groups. A mapping φ from set M to set $\hat{M}$ is said to be a homomorphism if i) for any elements m and m' of M, $\varphi(m - m') = \varphi(m) - \varphi(m')$, and ii) for any 0-neighborhood $\hat{U}$ in $\hat{M}$, $\varphi^{-1}[\hat{U}]$ is a 0-neighborhood in M. The first condition requires that the algebraic structure be preserved by the mapping. [Note that, in the equation of that condition, the "-" on the left is in M, that on the right in $\hat{M}$.] Setting $m = m' = 0$ in that equation, we obtain $\varphi(0) = \varphi(0) - \varphi(0)$. But the right side is zero (in $\hat{M}$) no matter what $\varphi(0)$ is: hence, by that equation, $\varphi(0) = 0$. Now set $m = 0$ in our equation, to obtain $\varphi(-m') = \varphi(0) - \varphi(m')$. But, since $\varphi(0) = 0$, we have $\varphi(-m') = -\varphi(m')$. Finally, setting $m'' = -m'$ in our equation, and using what we have just shown, we obtain $\varphi(m + m'') = \varphi(m) + \varphi(m'')$. In short, the single equation in the first condition of the definition requires that φ preserve sums, inverses, and zero (and conversely, as is easily checked, these three properties imply the equation of the first condition). The second condition requires that "you tell me how close you want $\varphi(m)$ to be to zero in $\hat{M}$, and I'll tell you how close m must be to zero in M".

Example. Let M and $\hat{M}$ be the reals. Fix any number a, and let φ have action $\varphi(r) = ar$. This is a homomorphism. Indeed, the first condition follows from the fact that $a(r - r') = ar - ar'$ for any numbers r and r'. For the second condition, consider 0-neighborhood $\hat{U} = (-\epsilon, \epsilon)$ in $\hat{M}$. Then $\varphi^{-1}[\hat{U}]$ consists of all r in M with ar in $(-\epsilon, \epsilon)$. But this is just $(-\epsilon/a, \epsilon/a)$ if a is positive, $(\epsilon/a, -\epsilon/a)$ if a is negative, and M if a is zero. In each case, we obtain a 0-neighborhood in M. Note in particular that, for $a = 0$, φ sends every element of M to zero in $\hat{M}$.

Example. Let M be $\mathbb{R}^n$ (second example on page 7), and $\hat{M}$ the reals. Let φ be the mapping with action $\varphi(r_1, \ldots, r_n) = r_1$, i.e., φ "takes

the first component". Then φ is a homomorphism. For the first condition, $\varphi((r_1,\ldots,r_n) - (r_1',\ldots,r_n')) = \varphi(r_1-r_1',\ldots,r_n-r_n')$ $= r_1 - r_1' = \varphi(r_1,\ldots,r_n) - \varphi(r_1',\ldots,r_n')$, where we used the definition of φ in the second and third steps. For the second condition, let $\hat{U} = (-\epsilon, \epsilon)$ be a 0-neighborhood in $\hat{M}$. Then $\varphi^{-1}[\hat{U}]$ consists of all $(r_1,\ldots,r_n)$ with r_1 in $(-\epsilon,\epsilon)$. But this is a 0-neighborhood in M, for it contains as a subset the collection of all $(r_1,\ldots,r_n)$ in M with $[(r_1)^2 + \ldots + (r_n)^2]^{1/2} < \epsilon$.



Example. Let M be the reals, and $\hat{M}$ $\mathbb{R}^n$. Let φ be the mapping with action $\varphi(r)$ $= (r,0,\ldots,0)$. Then φ is a homomorphism.

Example. Let M be the reals, and let $\hat{M}$ be R^2. Fix a number $a \neq 0$ and let φ have action $\varphi(r) = (r,a)$. This is not a homomorphism, for $\varphi(r - r') = (r-r',a)$, while $\varphi(r) - \varphi(r') = (r,a) - (r',a) = (r-r',0)$.

It is part of the definition that a homomorphism preserves sums of two (and therefore of any finite number) of group elements. It turns out, however, that "infinite sums" are also preserved. We claim: let $\Sigma m_i = m$, with m finite in M. Then $\Sigma \varphi(m_i) = \varphi(m)$. Proof: Let $\varphi(m) + \hat{U}$ be any $\varphi(m)$-neighborhood in $\hat{M}$. Then $m + \varphi^{-1}[\hat{U}]$ is an m-neighborhood in M. By convergence, there is a finite set K of positive integers such that, for any larger finite K', $\sum_{K'} m_i$ is in $m + \varphi^{-1}[\hat{U}]$. Hence, $\varphi(\sum_{K'} m_i)$ is in $\varphi[m + \varphi^{-1}[\hat{U}]]$. But, by the first condition for a homomorphism, the first is $\sum_{K'} \varphi(m_i)$, and the second $\varphi(m) + \hat{U}$. We have shown that, for any 0-neighborhood $\hat{U}$ in $\hat{M}$, there is a finite K such that, for any larger finite K', $\sum_{K'} \varphi(m_i)$ is in $\varphi(m) + \hat{U}$, which is precisely what was to be proven.

Note that convergence of the original series above was necessary:

Example. Set n = 2 in the second example on the previous page. Set $m_1 = m_2 = \ldots = (0,1)$. Then $\Sigma m_i = \infty$. But $\varphi(m_i) = 0$ for each i, whence $\Sigma \varphi(m_i) = 0$.

There is nothing very suprising here: if all the structure needed to sum a series is preserved, then sums of series are preserved.

13. Changing the Group

We now return to our program of constructing measure spaces from measure spaces.

Let $X, \mathcal{M}, M, \mu$ be a measure space, let $\hat{M}$ be a complete abelian topological group, and let φ be a mapping from set M to set $\hat{M}$. We wish to construct, in the presence of this set-up, a measure space based on group $\hat{M}$. Clearly, we can do nothing unless we first assume more about φ, namely, that it interacts in some appropriate way with the structures on M and $\hat{M}$. We therefore suppose, as the discussion of the previous section suggests, that φ is in fact a homomorphism. The idea, then, would be the following. We keep the same measurable space, $X, \mathcal{M}$. We replace M by $\hat{M}$. We now wish to define $\hat{\mu}$: it is to assign, to each measurable set, some element of $\hat{M}^*$. The obvious choice would be this: if $\mu(A)$ is finite, set $\hat{\mu}(A) = \varphi(\mu(A))$ (i.e., one finds the μ-measure of A, and sends that element of M to $\hat{M}$ via φ), while if $\mu(A)$ is infinite, set $\hat{\mu}(A) = \infty$. Unfortunately, we do not in this way obtain a measure space in general:

Example. Let $X, \mathcal{M}, M, \mu$ be Lebesque measure, let $\hat{M}$ be the complete abelian topological group with the single element "0", and let φ be the homomorphism with action $\varphi(m) = 0$ for every m in M. Let $\hat{\mu}$ be that defined above. Then, for any set A of finite Lebesque measure, $\hat{\mu}(A) = 0$, while, for any A of infinite Lebesque measure, $\hat{\mu}(A) = \infty$. But this is not a measure space (for, e.g., X has a partition by sets of finite Lebesque measure).

It is the usual difficulty, and, as usual, the cure is Theorem 10.

Let $X, \mathcal{M}, M, \mu$ be a measure space, $\hat{M}$ a complete abelian topological group, and φ a homomorphism from M to $\hat{M}$. Denote by $\mathcal{F}$ the collection of all measurable sets with $\mu(A)$ finite. For A in $\mathcal{F}$, set $\nu(A) = \varphi(\mu(A))$. We now claim that $X, \mathcal{M}, \mathcal{F}, \hat{M}, \nu$ is a pre-measure space. Proof: The first three conditions for a pre-measure space are immediate from the example on page 32. For the fourth condition, let $\mu(A)$ be finite, and let $A_1, A_2, \ldots$ be a partition of A. Then $\Sigma \mu(A_i) = \mu(A)$ by Theorem 6. Hence, from the result on page 41, $\Sigma \varphi(\mu(A_i)) = \varphi(\mu(A))$. That is, $\Sigma \nu(A_i) = \nu(A)$.

Thus, we obtain a pre-measure space, and so, by Theorem 10, a measure space, $X, \mathcal{M}, \hat{M}, \hat{\mu}$.

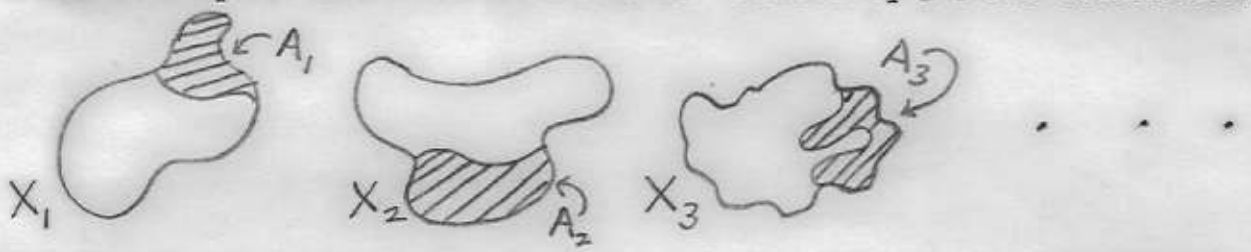
Example. Let $X, \mathcal{M}, M, \mu$ be any measure space, $\hat{M}$ the group with its only element "0", and φ with action $\varphi(m) = 0$. Then $X, \mathcal{M}, \hat{M}, \hat{\mu}$ is the following: $\hat{\mu}(A) = 0$ if A is μ-sigma-finite, and ∞ otherwise.
Example. Let $X, \mathcal{M}, M, \mu$ be Lebesque measure, let $\hat{M}$ be the reals, and let φ be that given in the first example on page 40. Thus, any "numerical multiple" of Lebesque measure is a measure space.

What properties of measure spaces are preserved under the construction above? First, we claim: $X, \mathcal{M}, \hat{M}, \hat{\mu}$ is sigma-finite if and only if $X, \mathcal{M}, M, \mu$ is. This is immediate from the following facts: every set with μ-finite measure has $\hat{\mu}$-finite measure, and every set with $\hat{\mu}$-finite measure has a partition by sets of μ-finite measure. Similarly, if $X, \mathcal{M}, M, \mu$ is finite, so is $X, \mathcal{M}, M, \mu$ (although the converse is false, by the first example above). This example also shows that a complete measure space may, by our construction, become an incomplete one.

14. Disjoint Unions

We now consider measure spaces constructed by "putting several measure spaces together".

Let $(X_1, \mathcal{M}_1, M, \mu_1)$, $(X_2, \mathcal{M}_2, M, \mu_2)$ . . . be a sequence of measure spaces (all with the same group M). We may take the underlying sets, $X_1, X_2, \ldots$, to be disjoint, for, were they not, we could replace each X_i by a copy of this set so that they would be disjoint. Let the X's, then, be disjoint. Set $X = \cup X_i$. We wish to construct a measure space based on this set X. Denote by $\mathcal{M}$ the collection of



all subsets A of X of the form $A = A_1 \cup A_2 \cup A_3 \cup \ldots$, where each A_i is a measurable subset of X_i. Then, as is easily checked, this $X, \mathcal{M}$ is a measurable space.

What one would like to do is set, for such an A, $\mu(A) = \Sigma \mu_i(A_i)$, i.e., to let the measure of A be the sum of the measures of its parts in each X_i. The usual type of example shows, however, that this need not lead to a measure space because of sets with infinite measure. We proceed therefore, as follows. Denote by $\mathcal{F}$ the collection of all measurable sets in X of the form $A = A_1 \cup A_2 \cup \ldots \cup A_n$ (note now only a finite number), with each of $\mu_1(A_1), \ldots, \mu_n(A_n)$ finite. For such an A, set $\nu(A) = \mu_1(A_1) + \ldots + \mu_n(A_n)$. Then $X, \mathcal{M}, \mathcal{F}, M, \nu$ is a pre-measure space. By Theorem 10, we obtain a measure space $X, \mathcal{M}, M, \mu$, called the disjoint union of measure spaces $(X_1, \mathcal{M}_1, M, \mu_1)$, $(X_2, \mathcal{M}_2, M, \mu_2)$, . . .

Example. Let X_1 have single element x_1, and let $\mu_1(x_1) = m_1$; let X_2 have single element x_2, and let $\mu_2(x_2) = m_2$; etc. The disjoint union of these measure spaces is precisely the measure space in the example on page 15.

Example. Let $X, \mathcal{M}, M, \mu$ be any measure space, and let $A_1, A_2, \ldots$ be a partition of X. Then each A_i defines a measure subspace. The disjoint union of these measure spaces is just the original measure space $X, \mathcal{M}, M, \mu$.

15. Sums of Measures; Spaces of Measures

Let $X, \mathcal{M}$ be a measurable space, and M a complete abelian topological group. A mapping μ from $\mathcal{M}$ to M* such that $X, \mathcal{M}, M, \mu$ is a measure space is called a measure (on $X, \mathcal{M}$, valued in M*). In this section, we do two things. First, we introduce algebraic operations (namely, those available in an abelian group: sums, inverses, a zero) on measures. Second, we introduce the space of measures (for fixed $X, \mathcal{M}$, M), and study what structure is available on this space.

Fix, once and for all, a measurable space $X, \mathcal{M}$, and a complete abelian topological group M. Let μ_1 and μ_2 be measures. We wish to define their sum. To this end, denote by $\mathcal{F}$ the collection of all measurable sets A with $\mu_1(A)$ and $\mu_2(A)$ finite, and, for such A, set $\nu(A) = \mu_1(A) + \mu_2(A)$. Then $X, \mathcal{F}, \mathcal{M}, \mathcal{F}, \nu$ is a pre-measure space. The measure μ which results from Theorem 10 is called the sum of measures μ_1 and μ_2, and is written $\mu = \mu_1 + \mu_2$. Note that $\mu = \mu_1 + \mu_2$ does not mean that $\mu(A) = \mu_1(A) + \mu_2(A)$ for every measurable A:

Example. Let X be a countable set; all subsets of X measurable; M the reals. For A any subset of X, let $\mu_1(A)$ be the number of elements of A (or infinity, if A has an infinite number of elements), and let $\mu_2(A)$ be minus that number (or infinity). Then $(\mu_1(X) + \mu_2(X) = \infty + \infty = \infty$, while $(\mu_1 + \mu_2)(X) = 0$.
This example will be recognized as our usual one, and also as the reason why, above, we made an excursion through Theorem 10. Of course, the sum of two sigma-finite (resp., finite) measures is sigma-finite (resp., finite).

Since this operation is called "sum", it is natural to see which, if any, of the conditions for an abelian group are satisfied by it. We claim, first, that $\mu_1 + \mu_2 = \mu_2 + \mu_1$. Proof: Since M is an abelian group, the pre-measure space used to obtain $\mu_1 + \mu_2$ is precisely the same as the pre-measure space used to obtain $\mu_2 + \mu_1$. We claim, furthermore, that $(\mu_1 + \mu_2) + \mu_3 = \mu_1 + (\mu_2 + \mu_3)$. Proof: Denote by $\mathcal{F}$ the collection of all measurable sets A with $\mu_1(A), \mu_2(A)$, and $\mu_3(A)$ all finite. For such A, set $\nu(A) = \mu_1(A) + \mu_2(A) + \mu_3(A)$. Then $X, \mathcal{M}, \mathcal{F}, M, \nu$ is a pre-measure space. But this pre-measure space is easily seen to give rise to both the

measure $(\mu_1 + \mu_2) + \mu_3$ and the measure $\mu_1 + (\mu_2 + \mu_3)$. Hence, these two measures must be equal. Thus, we have so far defined addition of measures, and have verified the first two conditions for an abelian group. We come next to the third condition: we must find some measure μ_0 such that $\mu + \mu_0 = \mu$ for every measure μ. Let μ_0 have action $\mu_0(A) = 0$ for every measurable A (obviously a measure); this μ_0 is called the zero measure, and is normally written "0". We now claim that, for any measure μ, $\mu + 0 = \mu$. Proof: In the pre-measure space used to define the sum on the left, $\mathcal{F}$ consists of all A with $\mu(A)$ finite (since $0(A) = 0$ is always finite), and, for such an A, $\nu(A) = \mu(A)$. By the remark just after the proof on page 34, $\mu + 0 = \mu$. Thus, the first three conditions for an abelian group are indeed satisfied.

Thus, the first three conditions for an abelian group are indeed satisfied. We consider, finally, the fourth. Let μ be any measure, and let $(-\mu)$ have the following action: for any measurable A, $(-\mu)(A) = -\mu(A)$. That is, $(-\mu)(A)$ is the inverse of the element $\mu(A)$ of M if $\mu(A)$ is finite, and $(-\mu)(A)$ is infinite if $\mu(A)$ is infinite. Then $-\mu$ is certainly a measure (e.g., by inspection, or by Section 13, noting that the mapping from M to M which sends m to -m is a homomorphism). This $-\mu$ is a good candidate for the inverse of μ, whose existence is required by the fourth condition for an abelian group. Unfortunately,

Example. Let X be any non-empty set; all subsets of X measurable. Let $\mu(A) = 0$ if $A = \emptyset$, and $\mu(A) = \infty$ if A is not empty. Then μ is a measure, and, clearly, $-\mu = \mu$ (since $-0 = 0$ in M). We find the measure $\mu + (-\mu)$. In the corresponding pre-measure space, $\mathcal{F}$ consists only of $\emptyset$, and $\nu(\emptyset) = 0$. Applying Theorem 10, we see that every non-empty set is in class iv) therein. Hence, $\mu + (-\mu) = \mu$. But $\mu \neq 0$, and so we do not have $\mu + (-\mu) = 0$.

Thus, the fourth condition for an abelian group fails. The situations under which it fails, as suggested by the above example, are these. We claim: $\mu + (-\mu) = 0$ if and only if μ is sigma-finite. Proof: Suppose first that μ is essentially infinite, so there is some μ-essentially infinite measurable set A. Then A is also $(-\mu)$-essentially infinite. Then, in the application of Theorem 10, this A will be in class iv), whence $(\mu + (-\mu))(A) = \infty$. Thus, we cannot

have $\mu + (-\mu) = 0$, for the measures on each side differ in their action on this A. Now suppose, conversely, that μ is sigma-finite. Then also $-\mu$ is sigma-finite. Then, in our pre-measure space, $\mathcal{F}$ will consist of all A with $\mu(A)$ finite (since in this case $(-\mu)(A)$ will necessarily be finite), and, for such A, we shall have $\nu(A) = 0$. We now apply Theorem 10. Since μ and $(-\mu)$ are both sigma-finite, no A will be in class iv) of that theorem. Since $\nu(A_i) = 0$ for every A_i in $\mathcal{F}$, no A will be in class iii). Thus, every A will be in class i) or ii), whence $(\mu + (-\mu))(A)$ will be zero for every A. That is, we have $\mu + (-\mu) = 0$.

Denote by $SF(X, \mathcal{M}, M)$ the collection of all sigma-finite measures on $X, \mathcal{M}$, valued in M. We have shown above that, under addition of measures, $SF(X, \mathcal{M}, M)$ is an abelian group.

In more detail, the mechanism for the failure of the fourth condition (for an abelian group) in general is this. Let μ_1 and μ_2 be measures, and consider $\mu_1 + \mu_2$. Then a set is $(\mu_1 + \mu_2)$-essentially infinite if and only if it is either μ_1-essentially infinite or μ_2-essentially infinite. Thus, under addition of measures, "the number of sets which are essentially infinite can only get larger". Thus, there is no way to have inverses of essentially infinite measures, for the addition of anything to such a measure will always leave at least some sets essentially infinite. The problem, then, is that one keeps accumulating essentially infinite sets, and there is no way to get rid of them by addition. If one restricts consideration to sigma-finite measures (so no sets are essentially infinite), then the problem does not arise. One could, alternatively, proceed in a more general way, e.g., along the following lines. Fix a collection $\mathcal{S}$ of measurable sets, and consider the set of all measures for which the sets in $\mathcal{S}$ - and only those sets - are essentially infinite. The sum of two such measures will again be one. We introduce, in this collection, a "zero measure", $0_{\mathcal{S}}$, with the following action: $0_{\mathcal{S}}(A) = 0$ if A is not in $\mathcal{S}$, and $0_{\mathcal{S}}(A) = \infty$ if A is in $\mathcal{S}$. Inverses are defined as before. In this way, we would obtain an abelian group of all measures whose essentially infinite sets are precisely those in $\mathcal{S}$, where the "zero" in the group is $0_{\mathcal{S}}$. The special case when $\mathcal{S}$ happens to contain no sets would be just $SF(X, \mathcal{M}, M)$ above. This generalization perhaps sheds more confusion than light, and therefore we shall, instead,

simply assume sigma-finiteness when necessary.

Thus, we have an abelian group, $SF(X, \mathcal{M}, M)$. It is natural to ask whether or not we can make this group into an abelian topological group, i.e., whether or not we can find an appropriate collection of 0-neighborhoods in $SF(X, \mathcal{M}, M)$ satisfying the conditions on page 7. Thus, we wish to introduce the notion of a measures being "close to the zero measure". We proceed as follows. Let U be any 0-neighborhood in M. Denote by $\tilde{U}$ the collection of all μ in $SF(X, \mathcal{M}, M)$ such that $\mu(A)$ is in U for every measurable A. [Thus, e.g., for U "very small", $\tilde{U}$ contains only measures which assign "very small measure" to each measurable set.] Call a subset of $SF(X, \mathcal{M}, M)$ a 0-neighborhood if it contains some $\tilde{U}$ as a subset. We now claim that, with these 0-neighborhoods, $SF(X, \mathcal{M}, M)$ is an abelian topological group. Proof: We must verify the five conditions on page 7. 1. Since the zero measure is in every $\tilde{U}$, it is in every 0-neighborhood in $SF(X, \mathcal{M}, M)$. Conversely, let μ be in every 0-neighborhood in $SF(X, \mathcal{M}, M)$, i.e., let μ be in every $\tilde{U}$. Let A be any measurable set. Then $\mu(A)$ is in every 0-neighborhood U in M, whence $\mu(A) = 0$. Since $\mu(A) = 0$ for every A, $\mu = 0$. Thus, the only measure in every 0-neighborhood in $SF(X, \mathcal{M}, M)$ is the zero measure. 2. Let U and V be 0-neighborhoods in M: we claim, then, that $\tilde{U} \cap \tilde{V} = \widetilde{(U \cap V)}$. Indeed, μ is in $\tilde{U} \cap \tilde{V}$ if and only if μ is in both $\tilde{U}$ and $\tilde{V}$, i.e., if and only if, for every measurable A, $\mu(A)$ is in both $\tilde{U}$ and $\tilde{V}$, i.e., if and only if, for every A, $\mu(A)$ is in $U \cap V$, i.e., if and only if μ is in $\widetilde{U \cap V}$. The second condition is now immediate from the fact that the intersection of any two 0-neighborhoods in M is another. 3. Any subset of $SF(X, \mathcal{M}, M)$ which contains a 0-neighborhood contains some $\tilde{U}$, and hence is itself a 0-neighborhood. 4. We claim that, for any 0-neighborhood U in M, $-\tilde{U} = \widetilde{(-U)}$. Indeed, μ is in $(-\tilde{U})$ if and only if $-\mu$ is in $\tilde{U}$, i.e., if and only if $-\mu(A)$ is in U for every A, i.e., if and only if $\mu(A)$ is in $-U$ for every A, i.e., if and only if μ is in $\widetilde{(-U)}$. 5. We claim first that, for any 0-neighborhood V in M, $\tilde{V} + \tilde{V} \subset \widetilde{(V + V)}$. Indeed, for μ_1 and μ_2 each in $\tilde{V}$, both $\mu_1(A)$ and $\mu_2(A)$ are in V for each measurable A, whence $\mu_1(A) + \mu_2(A)$ is in $V + V$ for each A, whence $(\mu_1 + \mu_2)$ is in $\widetilde{(V + V)}$. Now consider some 0-neighborhood $\tilde{U}$ in $SF(X, \mathcal{M}, M)$. Choose 0-neighborhood V in M with $V + V \subset U$. Then $\tilde{V} + \tilde{V} \subset \widetilde{(V + V)} \subset \tilde{U}$. This completes the proof.

Thus, $SF(X, \mathcal{M}, M)$ is an abelian topological group. We next show that this group is in fact complete. To this end, let μ_1, μ_2, . . . be a Cauchy sequence in $SF(X, \mathcal{M}, M)$, so, for every 0-neighborhood U in M, there is a number N such that $\mu_n - \mu_{n'}$ is in $\tilde{U}$ whenever $n \geq N$ and $n' \geq N$. We must find some element μ of $SF(X, \mathcal{M}, M)$ to which this sequence converges.

Consider first the 0-neighborhood M in M. Then, since our sequence is Cauchy, there is a number N such that $\mu_n - \mu_{n'}$ is in $\tilde{M}$ whenever $n \geq N$ and $n' \geq N$. In other words, whenever $n \geq N$ and $n' \geq N$, $(\mu_n - \mu_{n'})(A)$ is finite for every A. In still other words, whenever $n \geq N$ and $n' \geq N$, either both $\mu_n(A)$ and $\mu_{n'}(A)$ are finite, or both are infinite. Without loss of generality, let our Cauchy sequence begin with this N. Then we have: for any measurable A, either each $\mu_n(A)$ is finite, or each is infinite.

Next, denote by $\mathcal{F}$ the collection of all measurable sets A with all $\mu_n(A)$ finite. Then, since μ_1, μ_2, . . . is a Cauchy sequence in $SF(X, \mathcal{M}, M)$, $\mu_1(A)$, $\mu_2(A)$, . . . is a Cauchy sequence in M. Since M is complete, this Cauchy sequence converges to some element of M. Call this element $\nu(A)$, thus defining a mapping ν from $\mathcal{F}$ to M. We claim that $X, \mathcal{M}, \mathcal{F}, M, \nu$ is a pre-measure space. Proof: Fix a 0-neighborhood V, and a number N such that $\mu_n - \mu_{n'}$ is in $\tilde{V}$ whenever $n \geq N$ and $n' \geq N$. We proceed in steps as follows:

1. For A in $\mathcal{F}$, $\mu_N(A)$ is in $\mu(A) + V + V$. Since $\mu_n(A)$ converges to $\mu(A)$, there exists $n \geq N$ such that $\mu_n(A)$ is in $\mu(A) + V$. But $\mu_N(A)$ is in $\mu_n(A) + V$. Hence, $\mu_N(A)$ is in $\mu(A) + V + V$.

2. If A and B are disjoint sets in $\mathcal{F}$, then $\mu(A \cup B) = \mu(A) + \mu(B)$. By step 1, $\mu_N(A \cup B)$ is in $\mu(A \cup B) + V + V$, $\mu_N(A)$ is in $\mu(A) + V + V$, and $\mu_N(B)$ is in $\mu(B) + V + V$. From this, together with the fact that $\mu_N(A \cup B) = \mu_N(A) + \mu_N(B)$, we conclude that $\mu(A \cup B)$ is in $\mu(A) + \mu(B) + V + V + V + V - V - V$. Finally, since V is arbitrary, $\mu(A \cup B) = \mu(A) + \mu(B)$.

3. For A in $\mathcal{F}$, and A_1, A_2, . . . any partition of A, $\Sigma \mu(A_i) = \mu(A)$. We have $\Sigma \mu_N(A_i) = \mu_N(A)$. Hence, there is a finite K such that, for any finite K' containing K, $\sum_{K'} \mu_N(A_i)$ is in $\mu_N(A) + V$. But $\sum_{K'} \mu_N(A_i) = \mu_N(\bigcup_{K'} A_i)$; $\mu_N(\bigcup_{K'} A_i)$ is in $\mu(\bigcup_{K'} A_i) + V + V$, by step 1; $\mu(\bigcup_{K'} A_i) = \sum_{K'} \mu(A_i)$ by step 2; and $\mu_N(A)$ is in $\mu(A) + V + V$ by step 1. Putting these facts together, we have that $\sum_{K'} \mu(A_i)$ is in $\mu(A) + V + V + V - V - V$. Finally, since V is arbitrary, $\Sigma \mu(A_i) = \mu(A)$.

We now have that $X, \mathcal{M}, \mathcal{F}, M, \nu$ is a pre-measure space: the first three conditions follow from the fact that A is in $\mathcal{F}$ when and only when $\mu_1(A)$ is finite, and the fourth from step 3 above.

Now let μ be the measure which arises from this pre-measure space by Theorem 10. Then, since μ_1 is sigma-finite, and since $\mu(A)$ is finite whenever $\mu_1(A)$ is finite, we have that μ is sigma-finite. In other words, this μ is in $SF(X, \mathcal{M}, M)$. Finally, we note, from step 1 above, that the sequence $\mu_1, \mu_2, \ldots$ in $SF(X, \mathcal{M}, M)$ converges to μ.

To summarize, we have proven:

Theorem 11. Let $X, \mathcal{M}$ be a measurable space, and let M be a complete abelian topological group. Then $SF(X, \mathcal{M}, M)$, with addition and 0-neighborhoods defined as above, is a complete abelian topological group.

Example. Let X have single element x. Then $SF(X, \mathcal{M}, M) = M$ (for an element μ of $SF(X, \mathcal{M}, M)$ is characterized by $\mu(X)$).

Example. Let $X, \mathcal{M}, M, \mu$ be Lebesque measure. Set $\mu_1 = (1/2)\mu$, $\mu_2 = (3/4)\mu$, $\mu_3 = (7/8)\mu$, etc. Then the sequence $\mu_1, \mu_2, \ldots$ in $SF(X, \mathcal{M}, M)$ does not converge to μ (for, e.g., setting $U = (-1,1)$, there is no n with $\mu_n - \mu$ in $\tilde{U}$).

Example. Let $X, \mathcal{M}, M, \mu$ be a finite measure space. Fix measurable sets $A_1 \subset A_2 \subset \ldots$ with $\cup A_i = A$. For each i, let μ_i be the measure with action $\mu_i(B) = \mu(A_i \cap B)$. Then $\mu_1, \mu_2, \ldots$ converges to μ.

Example. Let $X, \mathcal{M}, M, \mu$ be Lebesque measure. Then, by Theorem 11, $SF(X, \mathcal{M}, M)$ is a compbete abelian topological group. For any measurable A, let $\hat{\mu}(A)$ be the following element of $SF(X, \mathcal{M}, M)$: $\hat{\mu}(A)$ is the measure on $X, \mathcal{M}$ which assigns to measurable B the number $\mu(A \cap B)$. Then $X, \mathcal{M}, SF(X, \mathcal{M}, M), \hat{\mu}$ is a measure space.

Note in particular that Theorem 11 provides one with a great variety of exotic and interesting complete abelian topological groups.

16. Direct Sums

As we have seen in Section 14, one can take the disjoint union of, e.g., two measure spaces X_1, $\mathcal{M}_1$, M_1, μ_1 and X_2, $\mathcal{M}_2$, M_2, μ_2, provided that $M_1 = M_2$. Consider, however, the situation when the groups are not equal. In this case, one would not know what the disjoint union is to mean, for the construction itself requires that one take sums such as $\mu_1(A_1) + \mu_2(A_2)$, while we do not know how to "add" elements from different groups. Similarly, in Section 15 we defined the sum $\mu_1 + \mu_2$ for measure spaces $X, \mathcal{M}, M_1, \mu_1$ and $X, \mathcal{M}, M_2, \mu_2$, provided $M_1 = M_2$. Again, the sum of measures makes no sense when M_1 and M_2 are different. In this section, we extend these two constructions to the case of different groups. The idea is to first introduce a way of "adding elements of different groups".

Let M_1 and M_2 be complete abelian topological groups. We construct, from these two groups, a new one. Denote by M the set consisting of all pairs, (m_1, m_2), with m_1 an element of M_1 and m_2 an element of M_2 (i.e., let $M = M_1 \times M_2$, the Cartesian product). We next introduce, on this set M, an operation of addition as follows: for (m_1, m_2) and (m_1', m_2') in M, set $(m_1, m_2) + (m_1', m_2') = (m_1 + m_1', m_2 + m_2')$ (i.e., the "+"s on the right are in M_1 and M_2; that on the left is being defined). We claim that this set M, with this operation of addition, is an abelian group. Proof: 1. $((m_1,m_2) + (m_1',m_2')) + (m_1'',m_2'') = (m_1+m_1',m_2+m_2') + (m_1'',m_2'') = (m_1+m_1'+m_1'', m_2+m_2'+m_2'') = (m_1,m_2) + ((m_1',m_2') + (m_1'',m_2''))$. 2. $(m_1,m_2) + (m_1',m_2') = (m_1+m_1',m_2+m_2') = (m_1',m_2') + (m_1,m_2)$. 3. For any (m_1,m_2) in M, $(m_1,m_2) + (0,0) = (m_1,m_2)$. Hence, the "0" in M is (0,0). 4. For any (m_1,m_2) in M, $(m_1,m_2) + (-m_1,-m_2) = (0,0)$. Hence, $-(m_1,m_2) = (-m_1,-m_2)$.

Thus, we obtain an abelian group M. We next wish to make this M into an abelian topological group, i.e., we wish to introduce 0-neighborhoods. We proceed as follows: a subset U of M will be called a 0-neighborhood if there exist 0-neighborhoods U_1 in M_1 and U_2 in M_2 such that (m_1, m_2) is in U whenever m_1 is in U_1 and m_2 in U_2. [That is, an element (m_1,m_2) of M is "near zero" if m_1 is near zero in M_1, and also m_2 is near zero in M_2.] We now claim that,

with these 0-neighborhoods, M is an abelian topological group. Proof: 1. (m_1,m_2) is in every 0-neighborhood in M if and only if m_1 is in every 0-neighborhood in M_1, and m_2 is in every 0-neighborhood in M_2, i.e., if and only if $(m_1,m_2) = (0,0)$. 2. If U contains all (m_1,m_2) with m_1 in U_1 and m_2 in U_2, and U' contains all (m_1,m_2) with m_1 in U_1' and m_2 in U_2', then $U \cap U'$ contains all (m_1,m_2) with m_1 in $U_1 \cap U_1'$ and m_2 in $U_2 \cap U_2'$. 3. Obvious. 4. If U contains all (m_1,m_2) with m_1 in U_1 and m_2 in U_2, then $-U$ contains all (m_1,m_2) with m_1 in $-U_1$ and m_2 in $-U_2$. 5. Let U contain all (m_1,m_2) with m_1 in U_1 and m_2 in U_2. Let $V_1 + V_1 \subset U_1$, and $V_2 + V_2 \subset U_2$. Then, if V consists of (m_1,m_2) with m_1 in V_1 and m_2 in V_2, $V + V \subset U$.

Thus, we have a complete abelian group M. Is M complete? Let (m_1^1,m_2^1), (m_1^2,m_2^2), . . . be a sequence in M. This sequence is Cauchy if and only if, for every 0-neighborhood U in M (say, U consisting of all (m_1,m_2) with m_1 in U_1 and m_2 in U_2), there is a number N such that (m_1^n,m_2^n) is in U whenever $n \geq N$. Clearly, then, this sequence is Cauchy if and only if the sequence $m_1^1, m_1^2, \ldots$ in M_1 is Cauchy, and the sequence $m_2^1, m_2^2, \ldots$ in M_2 is Cauchy. Furthermore, this sequence converges to (m_1, m_2) if and only if, for every 0-neighborhood U in M (say, consisting of all (m_1',m_2') with m_1' in U_1 and m_2' in U_2), there is a number N such that $(m_1^n, m_2^n) - (m_1, m_2)$ is in U whenever $n \geq N$. Clearly, then, this sequence converges to (m_1,m_2) if and only if the sequence $m_1^1, m_1^2, \ldots$ converges in M_1 to m_1, and the sequence $m_2^1, m_2^2, \ldots$ converges in M_2 to m_2. We conclude: since M_1 and M_2 are complete, so is M.

To summarize, given two complete abelian topological groups, M_1 and M_2, we obtain a third, M. This M is called the direct sum of M_1 and M_2, and is written $M = M_1 \oplus M_2$.

Example. Let each of M_1 and M_2 be the reals. Then M consists of all pairs, (r_1, r_2), of real numbers. Addition is "component-wise". That is, M is the abelian group $\mathbb{R}^2$. A subset U of M is a 0-neighborhood if and only if there exists positive numbers ϵ_1 and ϵ_2 such that U contains all (r_1, r_2) with $|r_1| \leq \epsilon_1$ and $|r_2| \leq \epsilon_2$. Clearly, then, U is a 0-neighborhood if and only if there is a positive ϵ such that (r_1, r_2) is in U whenever $[(r_1)^2 + (r_2)^2]^{1/2} \leq \epsilon$. In short, $M_1 \oplus M_2$ is precisely the first example on page 7.

In what sense is it true that the introduction of this M permits us to "add elements of different groups"? We may identify M_1 with the subset of M consisting of all $(m_1, 0)$, and M_2 with the subset consisting of all $(0, m_2)$. Then, since $(m_1,0) + (m_1',0) = (m_1 + m_1', 0)$, addition within M_1 is reflected by addition in M of those elements identified with elements of M_1, and similarly for M_2. Thus, for m_1 in M_1 and m_2 in M_2, although we don't know what $m_1 + m_2$ means, we do know what $(m_1,0) + (0,m_2)$ means in M, namely (m_1, m_2). In other words, to take the direct sum, one "combines the groups into a larger one, where that larger group is just large enough to contain a realization of each individual group, and such that it is a group (so addition can be performed therein)".

We can express the relationship between M_1, M_2, and $M_1 \oplus M_2$ in terms of mappings as follows. Let φ_1 be the mapping from M_1 to $M_1 \oplus M_2$ with action $\varphi(m_1) = (m_1,0)$. Then φ_1 is a homomorphism. Proof: 1. $\varphi_1(m_1) - \varphi_1(m_1') = (m_1,0) - (m_1',0) = (m_1-m_1',0)$ $= \varphi_1(m_1 - m_1')$. 2. Let U be a 0-neighborhood in M_1 say, containing all (m_1, m_2) with m_1 in U_1 and m_2 in U_2. Then $\varphi_1^{-1}[U]$ contains all m_1 in U_1, whence $\varphi_1^{-1}[U]$ is a 0-neighborhood in M_1. Similarly, the mapping φ_2 from M_2 to $M_1 \oplus M_2$ with action $\varphi_2(m_2) = (0,m_2)$ is a homomorphism. In a similar way, we obtain mappings from $M_1 \oplus M_2$ to each of the individual groups. Let γ_1 be the mapping from $M_1 \oplus M_2$ to M_1 with action $\gamma_1((m_1,m_2)) = m_1$. Then γ_1 is a homomorphism. Proof: 1. $\gamma_1((m_1,m_2)) - \gamma_1(m_1',m_2')) = m_1 - m_1' = \gamma_1((m_1-m_1',m_2-m_2'))$ $= \gamma_1((m_1,m_2) - (m_1',m_2'))$. 2. Let U_1 be a 0-neighborhood in M_1. Then $\gamma_1^{-1}[U_1]$ consists of all (m_1,m_2) with m_1 in U_1 and m_2 in M_2. But M_2 is itself a 0-neighborhood in M_2, and so $\gamma_1^{-1}[U_1]$ is a 0-neighborhood in $M_1 \oplus M_2$. Similarly, the mapping γ_2, with obvious action, from $M_1 \oplus M_2$ to M_2 is a homomorphism. Note that, e.g., $\gamma_1 \circ \varphi_1$ is the identity homomorphism on M_1. In the case when the groups are equal, $M_1 = M_2$, still another homomorphism is available. Let φ be the mapping from $M \oplus M$ to M with action $\varphi((m,m')) = m + m'$. Then φ is a homomorphism. Proof: 1. $\varphi(m,m') - \varphi(m'',m''') =$ $(m + m') - (m'' + m''') = (m - m'') + (m' - m''') = \varphi(m - m'',m' - m''')$. 2. Let U be a 0-neighborhood in M, and choose 0-neighborhood V with $V + V \subset U$. Then, for m and m' in V, $\varphi(m,m')$ is in U. Hence, $\varphi^{-1}[U]$ is a 0-neighborhood in $M \oplus M$.

To summarize, there are homomorphisms from each of the two summands into the direct sum, and homomorphisms from the direct sum to each of the two summands. For $M \oplus M$, "the enlargement of M to $M \oplus M$ to obtain addition is artificial, because we could just as well have added already within M", an observation which is reflected in a homomorphism from $M \oplus M$ back to M.

As regards sums of series, the situation is as follows. Let (m_1^1, m_2^1), (m_1^2, m_2^2), . . . be a series in $M_1 \oplus M_2$. Then we claim: $\Sigma(m_1^i, m_2^i) = (m_1, m_2)$ (finite) if $\Sigma m_1^i = m_1$ in M_1 and $\Sigma m_2^i = m_2$ in M_2, while $\Sigma(m_1^i, m_2^i) = \infty$ if either $\Sigma m_1^i = \infty$ or $\Sigma m_2^i = \infty$. Proof: From the observation on page 41, it follows that, if $\Sigma(m_1^i, m_2^i) = (m_1, m_2)$, finite, then $\Sigma m_1^i = m_1$ and $\Sigma m_2^i = m_2$. Suppose, conversely, that the sums in the individual groups converge. Let U be a 0-neighborhood in $M_1 \oplus M_2$, say, consisting of (m_1, m_2) with m_1 in U_1 and m_2 in U_2. Then there is a finite K_1 such that, for any finite K_1' containing K_1 as a subset, $\sum_{K_1'} m_1^i$ is in $m_1 + U_1$, and, similarly, a finite K_2. Let K consist of all i in either K_1 or K_2. Then, for K' finite, containing K as a subset, $\sum_{K'}(m_1^i, m_2^i)$ is in $(m_1, m_2) + U$. That is, the series in $M_1 \oplus M_2$ converges to (m_1, m_2).

We remark, finally, that the direct sum of more than two groups is constructed similarly, and its properties obtained similarly. [In fact, one can also take the direct product of an infinite collection of groups, although we shall not bother to do so.]

Example. Let X be a set with exactly n elements, let all subsets of X be measurable, and let M be any complete abelian topological group. Then $SF(X, \mathcal{M}, M) = M \oplus M \oplus \ldots \oplus M$ (n times).

Example. Let $X, \mathcal{M}$ be a measurable space, and M_1 and M_2 complete abelian topological groups. Then $SF(X, \mathcal{M}, M_1 \oplus M_2) = SF(X, \mathcal{M}, M_1) \oplus SF(X, \mathcal{M}, M_2)$.

These preliminaries out of the way, we now return to the question at hand: the definition of disjoint unions and sums for the case when the measures are valued in different groups.

Let $X_1, \mathcal{M}_1, M_1, \mu_1$ and $X_2, \mathcal{M}_2, M_2, \mu_2$ be measure spaces. Take X_1 and X_2 disjoint, and set $X = X_1 \cup X_2$. Denote by $\mathcal{M}$ the collection of all subsets of X of the form $A_1 \cup A_2$, with A_1 a measurable subset of X_1, and A_2 a measurable subset of X_2. Then, as we have seen in Section

14, $X, \mathcal{M}$ is a measurable space. Let $A = A_1 \cup A_2$ be a measurable set. Set $\mu(A) = (\mu_1(A_1), \mu_2(A_2))$ (in $M_1 \oplus M_2$) if $\mu_1(A_1)$ and $\mu_2(A_2)$ are finite, and $\mu(A) = \infty$ otherwise. Then $X, \mathcal{M}, M_1 \oplus M_2, \mu$ is a measure space. [Proof: Let $A = A_1 \cup A_2$. Choose a fine partition $A_1^1, A_1^2, \ldots$ of A_1 in X_1, and a fine partition $A_2^1, A_2^2, \ldots$ of A_2 in X_2. Then $A_1^i \cap A_2^j$, $(i,j = 1,2,\ldots)$ is a fine partition of A.] Our original construction of disjoint unions (when $M_1 = M_2$) can now be recovered from this construction as follows: consider the homomorphism from $M \oplus M$ to M, and use the construction of Section 13.

Similarly for sums. Let $X, \mathcal{M}, M_1, \mu_1$, and $X, \mathcal{M}, M_2, \mu_2$ be measure spaces. For any measurable A, set $\mu(A) = (\mu_1(A), \mu_2(A))$ if both $\mu_1(A)$ and $\mu_2(A)$ are finite, and $\mu(A) = \infty$ otherwise. Then, as above, $X, \mathcal{M}, M_1 \oplus M_2, \mu$ is a measure space. Again, when $M_1 = M_2$, the addition of measures in Section 15 is obtained from this construction via the homomorphism from $M \oplus M$ to M. Note that "addition of measures", as defined here, yields nothing like an abelian group, for, under addition, the complete abelian topological group M continues to get only larger. That is, there will be no inverses.

Example. Let $X, \mathcal{M}, \mathbb{R}^2, \mu$ be a measure space. Using the two homomorphisms from $\mathbb{R}^2$ to $\mathbb{R}$, we obtain two measure spaces, $X, \mathcal{M}, \mathbb{R}, \mu_1$ and $X, \mathcal{M}, \mathbb{R}, \mu_2$. Taking their sum, as in the paragraph above, we recover the original measure space $X, \mathcal{M}, \mathbb{R}^2, \mu$.

This last example is one of the reasons why "real measures" suffice for many applications. A "vector measure" (valued in $\mathbb{R}^2$) is equivalent to two "scalar measures" (i.e., the "components" of the original measure, valued in $\mathbb{R}$).

Example. Let $X, \mathcal{M}, M, \mu$ be a measure space. Let φ_1 be a one-to-one, onto mapping from X to X_1, and φ_2 a one-to-one, onto mapping from X to X_2. Using the method of Section 11, obtain measure spaces $X_1, \mathcal{M}_1, M, \mu_1$ and $X_2, \mathcal{M}_2, M, \mu_2$. Take their disjoint union as above, to obtain measure space $X_1 \cup X_2, \mathcal{N}, M \oplus M, \nu$. Now consider the mapping φ from $X_1 \cup X_2$ to X which sends x_1 in X_1 to x with $\varphi_1(x_1) = x$, and x_2 in X_2 to x with $\varphi_2(x_2) = x$. By Section 11, we obtain a measure space $X, \hat{\mathcal{M}}, M \oplus M, \hat{\mu}$. This measure space is precisely the sum above of μ and μ.

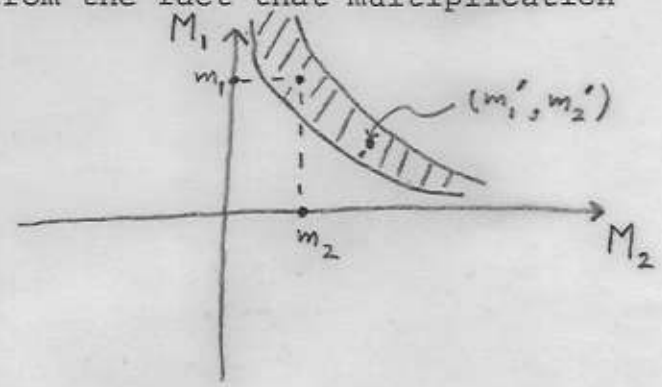
In short, sums can be constructed from disjoint unions, and vice versa.

17. Tensor Products

Let M_1 and M_2 be complete abelian topological groups. We saw in the previous section how one "adds" elements from these two groups: one creates a new complete abelian topological group, $M_1 \oplus M_2$, in which such sums are to lie. We shall also find it necessary to take "products" of one element from M_1 with one from M_2. Again, one makes available to himself this new operation by the introduction of a new group. We here construct this group.

It is convenient to first introduce the following definition. Let M_1, M_2, and G be complete abelian topological groups. Consider $M_1 \times M_2$, the Cartesian product of sets M_1 and M_2 (i.e., the set of all pairs, (m_1, m_2), with m_1 in M_1 and m_2 in M_2). A mapping φ from $M_1 \times M_2$ to G will be called a product mapping if i) for any elements m_1 and m_1' of M_1, and elements m_2 and m_2' of M_2, $\varphi(m_1+m_1', m_2) = \varphi(m_1, m_2) + \varphi(m_1', m_2)$ and $\varphi(m_1, m_2+m_2') = \varphi(m_1, m_2) + \varphi(m_1, m_2')$, and ii) given any m_1 in M_1 and m_2 in M_2, and any $\varphi(m_1,m_2)$-neighborhood $\varphi(m_1,m_2) + U$ in G, there is an m_1-neighborhood $m_1 + U_1$ and an m_2-neighborhood $m_2 + U_2$ such that $\varphi(m_1',m_2')$ is in $\varphi(m_1,m_2) + U$ whenever m_1' is in $m_1 + U_1$ and m_2' is in $m_2 + U_2$. One thinks of $\varphi(m_1, m_2)$ as the "product" of m_1 and m_2, where this product lies in G. Thus, the first condition above can be understood as requiring that the product be distributive. The second condition requires that "nearby elements (in M_1 and M_2) have nearby products", i.e., it requires that "the product is continuous".

Example. Let M_1, M_2, and G be the reals. For m_1 and m_2 real numbers, let $\varphi(m_1, m_2) = m_1 m_2$, the product of real numbers. This is a product mapping. The first condition follows from the fact that multiplication of numbers is distributive, i.e., $(m_1 - m_1')m_2 = m_1 m_2 - m_1' m_2$. For the second condition, fix numbers m_1 and m_2, and let ϵ be a positive number. Then the set of all numbers r with $|r - m_1 m_2| \leq \epsilon$ is an $m_1 m_2$-neighborhood in G. The set of all m_1' and m_2' with $|m_1' m_2' - m_1 m_2| \leq \epsilon$ is indicated in the figure above. Clearly, there are positive ϵ_1 and ϵ_2 such that, whenever $|m_1' - m_1| \leq \epsilon_1$ and $|m_2' - m_2| \leq \epsilon_2$, then $|m_1' m_2' - m_1 m_2| \leq \epsilon$.



The idea is now the following. We are given the groups M_1 and M_2, and we wish to find the group G and the product mapping φ. Of course, there will in general be many such (e.g., in the previous example, we could have chosen for G the group whose only element is 0, and for φ the mapping with action $\varphi(m_1,m_2) = 0$, also a product mapping). Thus, in order to get a unique G and a unique φ, we must require something further. The result we are after is this:

Theorem 12. Let M_1 and M_2 be complete abelian topological groups. Then there exists a complete abelian topological group G, together with a product mapping φ from $M_1 \times M_2$ to G, satisfying the following condition: given any other complete abelian topological group $\tilde{G}$, together with product mapping $\tilde{\varphi}$ from $M_1 \times M_2$ to $\tilde{G}$, there is a unique homomorphism γ from G to $\tilde{G}$ such that $\gamma \circ \varphi = \tilde{\varphi}$.

This Theorem will satisfy our needs for "taking products" in the following sense. First note that the G, φ whose existence is guaranteed by the theorem is unique. Indeed, let $\tilde{G}, \tilde{\varphi}$ be another. Then, by Theorem 12 for G, φ, there is a homomorphism γ from G to $\tilde{G}$ with $\gamma \circ \varphi = \tilde{\varphi}$, while, by Theorem 12 for $\tilde{G}, \tilde{\varphi}$, there is a homomorphism $\tilde{\gamma}$ from $\tilde{G}$ to G with $\tilde{\gamma} \circ \tilde{\varphi} = \varphi$. It then follows that $\tilde{\gamma} \circ \gamma$ is the identity on G, while $\gamma \circ \tilde{\gamma}$ is the identity on $\tilde{G}$. That is to say, the two homomorphisms γ and $\tilde{\gamma}$ "make G and $\tilde{G}$ identical as complete abelian topological groups". The G whose existence is guaranteed by Theorem 12 is called the tensor product of M_1 and M_2, written $M_1 \otimes M_2$. We now claim, furthermore, that the tensor product is "the most general product", in the following sense. Let G, φ be as in Theorem 12, and let $\tilde{G}$ be any complete abelian topological group, and γ any homomorphism from G to $\tilde{G}$. Then, as is easily checked, $\gamma \circ \varphi$ is a product mapping from $M_1 \times M_2$ to $\tilde{G}$. Thus, a homomorphism from one product (namely G), gives another. But Theorem 12 asserts that every product mapping from $M_1 \times M_2$ arises in precisely this way. Thus, once one has constructed the tensor product, all other product mappings from $M_1 \times M_2$ arise simply from homomorphisms from $M_1 \otimes M_2$ to other complete abelian topological groups. In short, Theorem 12 reduces the problem of finding product mappings from $M_1 \times M_2$ to that of finding homomorphisms from $M_1 \otimes M_2$.

The proof of Theorem 12, although essentially a straightforward construction of a G, φ that works, is long and rather tedious. We shall not actually prove Theorem 12 in detail, but, instead, will merely sketch the steps in the construction. There are five such steps: "getting a group", "reducing the group","getting some 0-neighborhoods", "reducing the group again", and "getting completeness".

1. Getting a group. Denote by S the set $M_1 \times M_2$. We introduce the following abelian group A (called the free abelian group on S). An element of the set A will be a formal, finite string of elements of S, joined by plus or minus signs, e.g., $- s_1 + s_2 + s_3 - s_4 - s_5$. Take two such strings as being equal in A if one can be obtained from the other by the operations of i) changing the order (so, e.g., the element above is equal to $+ s_2 - s_4 + s_3 - s_5 - s_1$), and ii) removing simultaneously "+s" and "-s" from the string (so, e.g., $s_1 + s_2 + s_3 - s_2 = s_1 + s_3$). [More precisely, A is a set of equivalence classes of "strings", where these operations define the equivalence relation.] We define addition in A by "running strings together", e.g., $(s_1 - s_2) + (- s_3 + s_4) = s_1 - s_2 - s_3 + s_4$. This set A, with this addition operation, is an abelian group (the zero in A is the string with no elements; for inverses, e.g., $- (-s_1 + s_2 - s_3) = + s_1 - s_2 + s_3$).

We next introduce a mapping α from $M_1 \times M_2$ to the abelian group A as follows. For $s = (m_1, m_2)$ in $M_1 \times M_2$, set $\alpha(s) = s$ (the element of A which is the "string" with this one element).

2. Reducing the Group. We now have a mapping α from $M_1 \times M_2$ to some abelian group A. We cannot ask whether or not this α is a product mapping, for A is only an abelian group, rather than an abelian topological group. We can, however, ask whether or not α satisfies the "algebraic condition" for a product mapping, i.e., whether or not $\alpha(m_1 - m_1', m_2) = \alpha(m_1, m_2) - \alpha(m_1', m_2)$, and similarly for M_2. The answer, however, is that α does not even satisfy this: let s_1, s_2, and s_3 be the elements $(m_1 - m_1', m_2)$, (m_1, m_2), and (m_1', m_2), respectively, of $M_1 \times M_2$. Then $\alpha(s_1) = s_1$, and $\alpha(s_2) - \alpha(s_3) = s_2 - s_3$. These two strings are not in general equal by our definition of equality in A.

The next step is to reduce the size of A to make α an "algebraic product mapping". Denote by K the collection of all elements of A which can be written as sums and differences of strings of the following

forms: $\alpha(m_1 - m_1', m_2) - \alpha(m_1, m_2) + \alpha(m_1', m_2)$ and $\alpha(m_1, m_2 - m_2') - \alpha(m_1, m_2) + \alpha(m_1, m_2')$. In short, K consists precisely of those elements of A which "ought to be zero, if α is to have the algebraic property of a product mapping". We note next that this K is a subgroup of A (i.e., K contains 0 in A; K contains the sum of any two elements in K; K contains the inverse of any element in K).

Set B = A/K, the quotient group. That is, the set B is the set of equivalence classes in A, where two elements of A are equivalent if their difference is in K. To add two elements of B, take representatives in A, add in A, and take the equivalence class in which that sum lies (to again get an element of B). This addition is independent of choice of representatives, because K is a subgroup.

We now define a mapping β from $M_1 \times M_2$ to B as follows. For (m_1, m_2) in $M_1 \times M_2$, let $\beta(m_1, m_2)$ be that element of B which is the equivalence class containing the element $\alpha(m_1, m_2)$ of A. In short, B results from A by "calling more things in A equal, namely things which differ by elements of A of the form of the first line of this page".

We now claim that the mapping β from $M_1 \times M_2$ to B satisfies the algebraic conditions for a product mapping, e.g., that $\beta(m_1 - m_1', m_2) = \beta(m_1, m_2) - \beta(m_1', m_2)$. To prove this, one just uses the fact that B was constructed by "setting equal to zero in A" these undesirable elements.

Example. Let M_1 and M_2 be the reals. Then A consists of formal strings of sums and differences of pairs of real numbers. Such strings as (3, -7) - (2, -7) - (1, -7), while nonzero in A, lie in the subgroup K, and hence are zero in B. [That is, this string lies in the equivalence class in A which also contains 0 in A.] We claim that, if $(r_1,r_2) + \ldots + (r_1',r_2')$ is zero in B, then $r_1r_2 + \ldots + r_1'r_2' = 0$ (products and sums of real numbers). Is the converse true, e.g., does $r_1r_2 = r_1'r_2'$ imply that $(r_1,r_2) = (r_1',r_2')$ in B? [If so, B would also be just the reals.] It is true, e.g., that (2,3) = (3,2) in B, for each is equal to (1,1) + (1,1) + (1,1) + (1,1) + (1,1) + (1,1). However, it is false, e.g., that $(\sqrt{2}, 1) = (1, \sqrt{2})$ in B.

3. Getting Some 0-Neighborhoods. We next wish to introduce, using the 0-neighborhoods on M_1 and M_2, some 0-neighborhoods on the abelian group B. It is convenient to have available, for purposes of

the present argument, the following definition. Call a system of 0-neighborhoods on B "acceptable" if it satisfies the following two conditions: i) the system of 0-neighborhoods on B satisfies all five conditions on page 7, but with the first condition replaced by: every 0-neighborhood contains 0. [That is to say, it is permitted that some element of B, different from 0, lie in every 0-neighborhood, something that is not permitted in an abelian topological group.], and ii) for any m_1 in M_1 and m_2 in M_2, and any 0-neighborhood U in our system, there exist 0-neighborhoods U_1 and U_2 in M_1 and M_2, respectively, such that $\beta(m_1', m_2')$ is in $\beta(m_1, m_2) + U$ whenever m_1' is in $m_1 + U_1$ and m_2' is in $m_2 + U_2$. [That is, we require of our system of 0-neighborhoods on B that, with respect to it, β satisfy the "neighborhood condition" for a product mapping.]

There will always be at least one acceptable system of 0-neighborhoods on B, e.g., that system whose only 0-neighborhood is B itself. [It was in order to be certain that there be at least one acceptable system that condition i) above was what it was.] In general, there will be many; we want the "largest". To this end, we proceed as follows. Denote by $\mathcal{U}$ the collection of all subsets U of B such that U is a 0-neighborhood in some acceptable system of 0-neighborhoods on B. Thus, the question of membership of some subset U of B in is a rather complicated business: it is the question of whether or not there is an entire system of acceptable 0-neighborhoods which includes U. One now verifies that $\mathcal{U}$ itself is an acceptable system of 0-neighborhoods (clearly, then, the "largest" acceptable system).

4. Reducing the Group Again. We have so far made the following progress. We have an abelian group B, and a system $\mathcal{U}$ of 0-neighborhoods on B. This B, $\mathcal{U}$ satisfies all the conditions for an abelian topological group, except possibly the first condition on page 7 (for this condition was weakened above). We have, furthermore, a mapping β from $M_1 \times M_2$ to B which satisfies the two conditions for a product mapping (i.e., which would be a product mapping, were B only a complete abelian topological group). The next step, then, is to "correct" B, to make it be an abelian topological group.

Denote by L the subset of B which is the intersection of all U in $\mathcal{U}$. This L is the thing which "ought to be zero in B, if B is to be an abelian topological group". One shows that this L is in fact a subgroup of B. [Quite easy: Since $\mathcal{U}$ is an acceptable system of 0-neighborhoods on B, it satisfies conditions 4 and 5 on page 7. That

L is indeed a subgroup is immediate from these two conditions.] Once again, we have managed to isolate the difficulty in a subgroup; once again, we take the quotient group. Let D be the abelian group $D = B/L$.

Not wanting to here undo what we just did in the previous step, we must now find 0-neighborhoods on this D. We proceed as follows. Let U be any subset of D. Each element of D is of course an equivalence class in B (namely, an equivalence class the difference of any two elements of which is in L). We denote by $\tilde{U}$ the subset of B which is the union in B of all the equivalence classes in U. That is to say, b in B is in $\tilde{U}$ provided the equivalence class containing B (some element of D) happens to be in U. We now call a subset U of D a 0-neighborhood in D if $\tilde{U}$ is a 0-neighborhood in B (i.e., if $\tilde{U}$ is in our acceptable system $\mathcal{U}$). Thus, we obtain an abelian group D, and some 0-neighborhoods on D. One now shows that this D, with these 0-neighborhoods, is in fact an abelian topological group, i.e., one verifies the five conditions on page 7. [That this should work out is perhaps not suprising. Indeed, B was already "practically an abelian topological group", lacking only the first condition on page 7. But L represents precisely the failure of this condition, and D is obtained by "setting to zero in B those elements which lie in L".]

We now define a mapping δ from $M_1 \times M_2$ to D as follows. For (m_1, m_2) in $M_1 \times M_2$, we let $\delta(m_1, m_2)$ be that element of D which is the equivalence class in B containing the element $\beta(m_1, m_2)$ of B. Thus, we now have an abelian topological group D, and a mapping δ from $M_1 \times M_2$ to D. One now verifies that this δ in fact satisfies the two conditions for a product mapping (again, not suprising, since this was already true for β and B, and D results from B by just "getting rid of undesirables").

Example. Consider the example on page 59. Then the element $(\sqrt{2}, 1) - (1, \sqrt{2})$ of B is in L. Indeed, the sequence (1,1), (1.4,1), (1.41,1), . . . in B converges, with respect to $\mathcal{U}$, to $(\sqrt{2}, 1)$, while this sequence is the same, in B, as (1,1), (1,1.4), (1,1.41), . . . , which converges to $(1,\sqrt{2})$. Thus, $(\sqrt{2},1) = (1,\sqrt{2})$ in D. Similarly, $(r_1,r_2) = (r_1',r_2')$ in D when and only when $r_1r_2 = r_1'r_2'$. Hence, D, in this example, is also the reals, and the mapping δ is just the product mapping of the example on page 56.

5. Getting Completeness. We are now nearly done with the construction. We have an abelian topological group D, and a product mapping β from $M_1 \times M_2$ to D. The only remaining problem is that this D may not be complete, as is required in the statement of Theorem 12. What we now wish to do, therefore, is "enlarge D to make it complete".

Denote by E the collection of all Cauchy sequences, $d_1, d_2, \ldots$, in D. Two such, $d_1, d_2, \ldots$ and $d_1', d_2', \ldots$, will be taken as equivalent if $d_1, d_1', d_2, d_2', \ldots$ is also a Cauchy sequence. [Thus, for example, two Cauchy sequences in D, which converge to the same element d of D, are equivalent.] One checks that this is indeed an equivalence relation. Denote by G the collection of all equivalence classes. We define addition within this set G as follows. Consider two elements of G, and take representatives $d_1, d_2, \ldots$ and $d_1', d_2', \ldots$ Then $d_1+d_1', d_2+d_2', \ldots$ is also a Cauchy sequence in D (as one checks). We now define the sum of our original two elements of G to be that element of G which, as an equivalence class of Cauchy sequences, contains the Cauchy sequence $d_1+d_1', d_2+d_2', \ldots$ One must check, of course, that this sum is well-defined, i.e., that the sum of two elements of G yields the same element of G no matter which representatives of the summands were chosen.

Thus, we have a set G, on which addition is defined. One now checks that this set, with this addition operation, is in fact an abelian group. [A representative of 0 in G is the Cauchy sequence $0, 0, \ldots$ in D; a representative of the inverse of the element of G with representative $d_1, d_2, \ldots$ is $-d_1, -d_2, \ldots$]

We next make the abelian group G into an abelian topological group as follows. Let U be a 0-neighborhood in D. Denote by $\tilde{U}$ the subset of G consisting of all elements of G of which every representative, $d_1, d_2, \ldots$, satisfies the following condition: there is some number N such that d_n is in U whenever $n \geq N$. That is, $\tilde{U}$ consists of elements of G whose representatives in D "eventually get within U of 0 in D". Now let the 0-neighborhoods in G consist of subsets of G which contain some $\tilde{U}$ as a subset. One checks that this abelian group G, with these 0-neighborhoods, is an fact an abelian topological group, i.e., that the five conditions on page 7 are indeed satisfied. One then checks, furthermore, that this (now) abelian topological group G is complete. [This is perhaps not suprising. D lacked only completeness, while G was so constructed that it provides a "place to converge to" for every

Cauchy sequence in D. Hence, one might expect that every Cauchy sequence in G will also have a "place to converge to".]

Thus, we have now obtained, from our abelian topological group D, a complete abelian topological group G. We next introduce a mapping φ from $M_1 \times M_2$ to G as follows. For (m_1, m_2) in $M_1 \times M_2$, let $\varphi(m_1, m_2)$ be that element of G having, as representative, the Cauchy sequence $\delta(m_1, m_2)$, $\delta(m_1, m_2)$, $\delta(m_1, m_2)$, . . . in D. That is to say, we "imbed D in G using constant Cauchy sequences, and use the mapping δ from $M_1 \times M_2$ to D to get a mapping φ from $M_1 \times M_2$ to G". Finally, one checks, using the fact that δ was a product mapping, that φ is also a product mapping.

Example. Suppose, in this last step, that D were the rationals. Then E would be the set of Cauchy sequences of rationals, and two such would be equivalent if they converged to the same real. Hence, G in this case would be the reals.

Example. If D were already complete, then, since every Cauchy sequence in D converges already in D, we would have G = D.

The G which results, from the construction above, from D is called the completion of D. [It was because of the existence of this construction that we did not regard the assumption of completeness at the beginning as very serious.]

Thus, the five steps above lead to a complete abelian topological group G, together with a product mapping φ from $M_1 \times M_2$ to G. This is our candidate for the G, φ of Theorem 12. All that remains, then, is to prove that this candidate indeed satisfies the condition of that theorem. This proof is along the following lines. Let $\tilde{G}$ be another complete abelian topological group, and $\tilde{\varphi}$ another product mapping. Let g be any element of G. We must find the homomorphism γ, i.e., we must find some element $\gamma(g)$ of $\tilde{G}$. One first shows (by going back through the construction of G in detail), that there is a sequence $g_1, g_2, \ldots$ in G which converges to g, and such that each g_i is of the form $\varphi(m_1,m_2) - \ldots + \varphi(m_1',m_2')$ (i.e., each g_i is a finite combination, in G, of elements that come from $M_1 \times M_2$ via φ). For each i (with g_i as above), let $\tilde{g}_i$ be the element $\tilde{\varphi}(m_1,m_2) - \ldots + \tilde{\varphi}(m_1',m_2')$ of $\tilde{G}$ (i.e., we do with $\tilde{\varphi}$ to get $\tilde{g}_i$ what we did with φ to get g_i). Thus, we obtain, from the Cauchy sequence $g_1, g_2, \ldots$ in G, a sequence $\tilde{g}_1, \tilde{g}_2, \ldots$ in $\tilde{G}$. One next shows that this sequence

in G is in fact Cauchy. Hence, since $\tilde{G}$ is complete, this Cauchy sequence converges to some element $\tilde{g}$ of $\tilde{G}$. Set $\gamma(g) = \tilde{g}$ (checking that this is well-defined, i.e., that one gets the same $\tilde{g}$ no matter what sequence $g_1, g_2, \ldots$ converging to g was chosen). In this way, one constructs one's mapping from G to $\tilde{G}$. One then checks, finally, that this γ is a homomorphism, that it is unique, and that it indeed satisfies $\gamma \circ \varphi = \tilde{\varphi}$.

In the discussion above, we have included all of the constructions, and have left out essentially all of the checks. Our claim, then, is that Theorem 12 is true, and that a proof is obtained by filling in all the missing checks above. For G, φ the tensor product of M_1 and M_2, one normally writes $m_1 \otimes m_2$ instead of $\varphi(m_1, m_2)$. We shall even allow ourselves to write this as $m_1 m_2$. That is to say, it is to be understood that such "products" live in the tensor product.

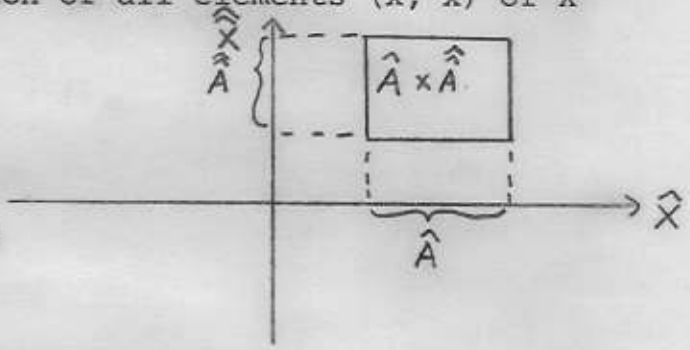
Example. The tensor product of the reals with the reals is again the reals (with the product mapping just multiplication of numbers). That is to say, given any product mapping $\tilde{\varphi}$ from $\mathbb{R} \times \mathbb{R}$ to $\tilde{G}$, where $\tilde{G}$ is any complete abelian topological group, there is some homomorphism γ from $\mathbb{R}$ to $\tilde{G}$ such that $\tilde{\varphi}(r_1, r_2) = \gamma(r_1 r_2)$. Explicitly, this homomorphism γ is, e.g., the following: for a any real number, set $\gamma(a) = \tilde{\varphi}(a,1)$. [$\gamma(a) = \tilde{\varphi}(1,a)$ would give the same homomorphism.]

In mathematics, theorems which are easy to state and easy to prove are normally just observations; those hard to state and hard to prove are normally highly computational lemmas; and those hard to state and easy to prove should normally never have been stated. It is a pleasure, occasionally, to run across a theorem which is easy to state and hard to prove.

18. Products of Measure Spaces

We come, finally, to our last construction which yields measure spaces from measure spaces: products.

Let $\hat{X}, \hat{\mathcal{M}}, \hat{M}, \hat{\mu}$ and $\hat{\hat{X}}, \hat{\hat{\mathcal{M}}}, \hat{\hat{M}}, \hat{\hat{\mu}}$ be measure spaces. Let X be the set $X = \hat{X} \times \hat{\hat{X}}$, i.e., the set of all pairs $(\hat{x}, \hat{\hat{x}})$ with $\hat{x}$ in $\hat{X}$ and $\hat{\hat{x}}$ in $\hat{\hat{X}}$. This X will be the underlying set of our new measure space. For $\hat{A}$ a measurable subset of $\hat{X}$, and $\hat{\hat{A}}$ a measurable subset of $\hat{\hat{X}}$, denote by $\hat{A} \times \hat{\hat{A}}$ the collection of all elements $(\hat{x}, \hat{\hat{x}})$ of X with $\hat{x}$ in $\hat{A}$ and $\hat{\hat{x}}$ in $\hat{\hat{A}}$. This $\hat{A} \times \hat{\hat{A}}$ will be called a rectangle in X. Note that the complement of a rectangle is the union of a finite collection of rectangles (in fact, of two, since $(\hat{A} \times \hat{\hat{A}})^c = \hat{A}^c \times \hat{\hat{X}} \cup \hat{X} \times \hat{\hat{A}}^c$), and that the intersection of any countable collection of rectangles is again a rectangle (since $\cap(\hat{A}_i \times \hat{\hat{A}}_i) = (\cap \hat{A}_i) \times (\cap \hat{\hat{A}}_i)$). In particular, the intersection of two rectangles is again a rectangle. Note, however, that the union of two rectangles need not be another.

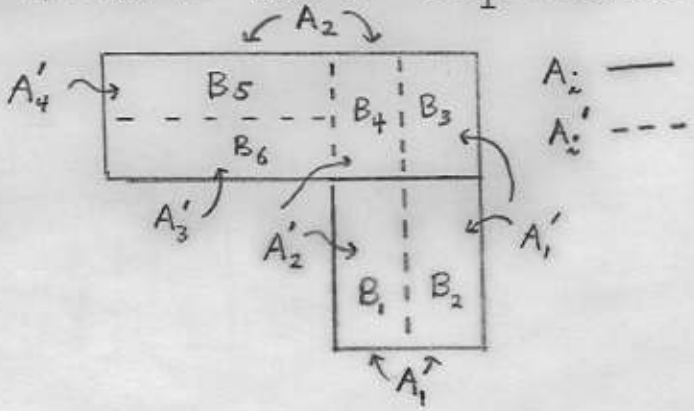


The idea is the following. Consider first a rectangle $A = \hat{A} \times \hat{\hat{A}}$, with both $\hat{\mu}(\hat{A})$ and $\hat{\hat{\mu}}(\hat{\hat{A}})$ finite. One would like to assign to such a rectangle an "area" which is "the product of its base ($\hat{\mu}(\hat{A})$) with its height ($\hat{\hat{\mu}}(\hat{\hat{A}})$)". The question is where this product is to live. One could imagine proceeding as follows. Fix some complete abelian topological group M, together with a product mapping φ from $\hat{M} \times \hat{\hat{M}}$ to M. One could then assign to our rectangle area $\varphi(\hat{\mu}(\hat{A}), \hat{\hat{\mu}}(\hat{\hat{A}}))$. Because φ is a product mapping, these assignments will satisfy "reasonable, product-like properties". One could indeed proceed along these lines, but it would be a bit awkward, for the following reason. In order to define the product of two measure spaces, one would need not only the measure spaces themselves, but, in addition, another group M together with some product mapping φ from $\hat{M} \times \hat{\hat{M}}$ to M. One would have many different products, corresponding to the many different choices of M, φ. A more natural procedure is suggested by the previous section. Consider Theorem 12. Suppose that we choose for M the tensor product, $\hat{M} \otimes \hat{\hat{M}}$, and for φ the product mapping from $\hat{M} \times \hat{\hat{M}}$ to $\hat{M} \otimes \hat{\hat{M}}$ given in that theorem. For these particular choices, we don't

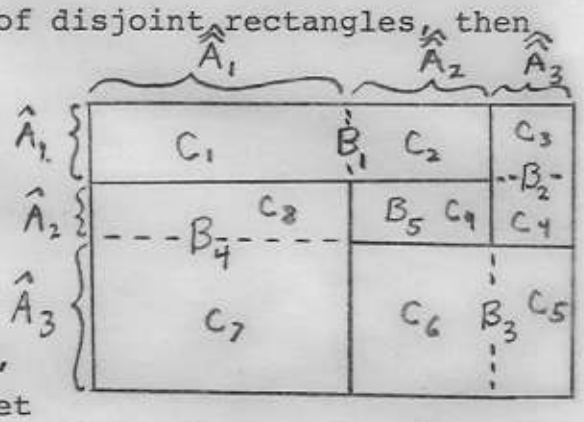
need any more information than that which was provided already in the original two measure spaces. There is, however, a more important reason to choose for M the tensor product, rather than some other group. Suppose that we have chosen $M = \hat{M} \otimes \hat{\hat{M}}$, and, with this choice, have constructed the product of our two measure spaces. Suppose that we then change our mind: we want to again take the product of our measure spaces, but with the measures in some other group M! (so we have some product mapping φ' from $\hat{M} \times \hat{\hat{M}}$ to M'). By Theorem 12, there is then some homomorphism γ from $\hat{M} \otimes \hat{\hat{M}}$ to M' such that $\varphi' = \gamma \circ \varphi$. But now, by Sect. 13 ("changing the group"), our measure valued in $\hat{M} \otimes \hat{\hat{M}}$, together with this homomorphism γ from $\hat{M} \otimes \hat{\hat{M}}$ to M', yields a measure valued in M'. Thus, we need not go through the entire construction of taking the product of our measure spaces again for M', φ': instead, we simply use this homomorphism γ. In short, because of the fundamental property of the tensor product of groups, we can choose the product measure to be in this tensor product, and need thereby take the product of measure spaces only once.

For $A = \hat{A} \times \hat{\hat{A}}$ a rectangle, with $\hat{\mu}(\hat{A})$ and $\hat{\hat{\mu}}(\hat{\hat{A}})$ finite, set $\lambda(A) = \hat{\mu}(\hat{A})\, \hat{\hat{\mu}}(\hat{\hat{A}})$ (an element of $\hat{M} \otimes \hat{\hat{M}}$). The idea is to extend the action of λ to more sets by taking unions of rectangles, etc.

We consider first finite unions. Let $A = A_1 \cup A_2 \cup \ldots \cup A_n$ be the union of a finite collection of rectangles (each with finite λ). Then, since the complement of a rectangle is the union of a finite number of rectangles, and since the intersection of a finite number of rectangles is a rectangle, we can always write this A as the union of a finite collection of disjoint rectangles. For $A = A_1 \cup \ldots \cup A_n$ with the A_i disjoint, set $\lambda(A) = \lambda(A_1) + \ldots + \lambda(A_n)$. We must, however, check that this is well-defined, i.e., that, if also $A = A_1' \cup \ldots \cup A_{n'}'$, disjoint, then $\sum_1^n \lambda(A_i) = \sum_1^{n'} \lambda(A_i')$. To this end, introduce nn' rectangles B_i, where each of these rectangles is of the form $A_i \cap A_j'$ ($i = 1,\ldots,n$; $j = 1,\ldots,n'$). Then $A = \cup B_i$, disjoint. It suffices to show that $\Sigma \lambda(A_i) = \Sigma \lambda(B_i)$. But note that each rectangle A_i is the union of a certain collection of the B-rectangles. Hence, we can rearrange the sum $\Sigma \lambda(B_i)$ to first sum over the rectangles in A_1,



then over those in A_2, etc. Suppose that it were true that the sum $\Sigma\,\lambda(B_i)$ over those B_i in A_1 were precisely $\lambda(A_1)$, and similarly for A_2, etc. Then, by rearrangement as above, we would indeed have $\Sigma\,\lambda(B_i) = \Sigma\lambda(A_i)$. Thus, we have only to show: if A is a rectangle, and $A = B_1 \cup B_2 \cup \cdots \cup B_n$, a union of disjoint rectangles, then $\lambda(A) = \Sigma\,\lambda(B_i)$. We prove this as follows. First "extend the edges of each B-rectangle across A", as in the figure. We thus obtain dissjoint rectangles C_i, such that A, as well as each of the B_j, is the union of various C-rectangles. Notwe, however, how the unions work in this case. Let $A = \hat{A} \times \hat{\hat{A}}$. Then there is a partition of $\hat{A}$, by $\hat{A}_1, \ldots, \hat{A}_{\hat{n}}$, and a partition of $\hat{\hat{A}}$, by $\hat{\hat{A}}_1, \ldots, \hat{\hat{A}}_{\hat{\hat{n}}}$, such that the C's are precisely the $\hat{n}\hat{\hat{n}}$ rectangles of the form $\hat{A}_i \times \hat{\hat{A}}_j$ $(i = 1,\ldots,\hat{n};\ j = 1,\ldots,\hat{\hat{n}})$. Similarly, for the decomposition of each B_j as the union of C's. Thus, we have only to prove: for $\hat{A}$, $\hat{\hat{A}}$ and $\hat{A}_1, \ldots, \hat{A}_{\hat{n}}$ and $\hat{\hat{A}}_1, \ldots, \hat{\hat{A}}_{\hat{\hat{n}}}$ as above, $\sum_{i=1}^{\hat{n}} \sum_{j=1}^{\hat{\hat{n}}} \lambda(\hat{A}_i \times \hat{\hat{A}}_j) = \lambda(\hat{A} \times \hat{\hat{A}})$ (for this would imply that $\Sigma\,\lambda(C_i) = \lambda(A)$, and also, by rearranging terms on the left, $\Sigma\,\lambda(C_i) = \Sigma\lambda(B_i)$, whence we shall have $\Sigma\lambda(B_i) = \lambda(A)$.) But it is indeed true that $\lambda(\hat{A} \times \hat{\hat{A}}) = \Sigma\,\Sigma\,\lambda(\hat{A}_i \times \hat{\hat{A}}_j)$, for the left side is $\hat{\mu}(\hat{A})\,\hat{\hat{\mu}}(\hat{\hat{A}})$ while the right side is $\Sigma\,\Sigma\,\hat{\mu}(\hat{A}_i)\,\hat{\hat{\mu}}(\hat{\hat{A}}_j)$, and equality holds because the mapping from $\hat{M} \times \hat{\hat{M}}$ to $\hat{M} \otimes \hat{\hat{M}}$ is a product mapping (i.e., is distributive). We conclude, then, that we indeed have a well-defined element $\lambda(A)$ of $\hat{M} \otimes \hat{\hat{M}}$ for A any finite union of rectangles.

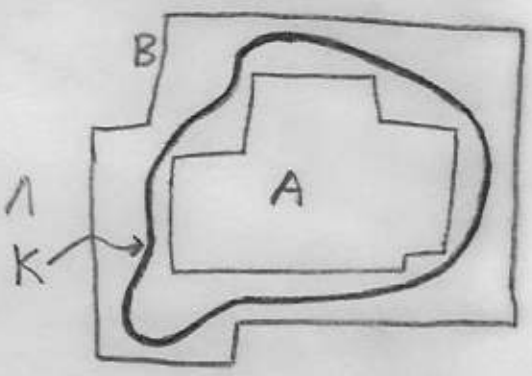


To summarize, for A and B each the union of a finite collection of rectangles (where each rectangle has finite base and height), $A \cup B$, $A \cap B$, and $A \cap B^c$ are also such. Furthermore, for A and B disjoint, $\lambda(A \cup B) = \lambda(A) + \lambda(B)$. [That is, we would have essentially a measure space, except that we cannot yet take countable unions.]

To get still more "measurable sets in the product", we proceed as follows. Denote by $\mathcal{F}'$ the [illegible] countable unions using rectangles. Then $X, \mathcal{M}$ is a measurable space - and, indeed, is the measurable space generated by rectangles on X. We want to construct a pre-measure space, and for this we need the collection $\mathcal{F}$ of "sets of finite measure". Just "finite unions of rectangles" will not do for $\mathcal{F}$, for it is not true in general that

every measurable subset of a finite union of rectangles is a finite union of rectangles. We must get more sets into our .

To this end, we proceed as follows. Consider a measurable subset K of X, and element m of $\hat{M} \otimes \hat{\hat{M}}$ with the following property: given any 0-neighborhood U in $\hat{M} \otimes \hat{\hat{M}}$, there are sets A and B, each a finite union of rectangles, such that $A \subset K \subset B$, and such that $\lambda(C)$ is in m + U whenever C is a finite union of rectangles with $A \subset C \subset B$. That is to say, "K can be approximated from the inside (by A) and from the outside (by B) by finite unions of rectangles, such that any finite union of rectangles (C) between A and B has $\lambda(C)$ near m". For K and m as above, the m is clearly unique. We denote by $\mathcal{F}'$ the collection of all such K's, and, for K in $\mathcal{F}'$, we set $\lambda(K) = m$. Thus, for example, any finite union K of rectangles is in $\mathcal{F}'$: set A = B = K.



Note that, for K and K' in $\mathcal{F}'$, so is $K \cup K'$, $K \cap K'$, and $K \cap K'^{c}$. [For example, let K and K' be in $\mathcal{F}'$. Find A, B, A', B', finite unions of rectangles, with $A \subset K \subset B$ and $A' \subset K' \subset B'$, as above. Now consider $A \cup A' \subset K \cup K' \subset B \cup B'$.] Furthermore, for K and K' disjoint in $\mathcal{F}'$, $\lambda(K \cup K') = \lambda(K) + \lambda(K')$ (which follows immediately from the corresponding fact for finite unions of rectangles).

We now have our candidate for a pre-measure space: $X, \mathcal{M}, \hat{M} \otimes \hat{\hat{M}}, \mathcal{F}', \lambda$. Indeed, conditions 1 and 3 on page 32 have already been proven. To prove condition 2, we must show that every measurable subset of a set in $\mathcal{F}'$ is itself in $\mathcal{F}'$. The essential step in proving this would, in turn, be to prove: given any rectangle (with finite base and height), and a subset of that rectangle which is a countable union of rectangles, that subset is necessarily in $\mathcal{F}'$. Similarly, the essential step in proving condition 4 on page 32 would be to prove this: given a rectangle A (with finite base and height), and given a partition A_1, A_2, . . . of A into disjoint rectangles, $\Sigma \lambda(A_i) = \lambda(A)$. The sad fact is, however, that, although both of these "essential steps" look reasonable, I know of neither a proof nor counterexample for either. It is easy to show in special cases, e.g., for Lebesque x Lebesque, that $X, \mathcal{M}, \hat{M} \otimes \hat{\hat{M}}, \mathcal{F}', \lambda$ is indeed a pre-measure space, and it is certainly

possible that this always holds.

Fortunately, our framework is sufficiently flexible to still salvage a "reasonable answer". Denote by $\mathcal{F}$ the collection of all elements K of $\mathcal{F}'$ satisfying: i) every measurable subset of K is in $\mathcal{F}'$, and ii) for any partition $K_1, \ldots$ of K, $\Sigma \lambda(K_i) = \lambda(K)$. Then, clearly, $X, \mathcal{M}, \hat{M} \otimes \hat{\hat{M}}, \mathcal{F}, \lambda$ is a pre-measure space, from which we may obtain a measure space. In this way, at least, we can take products of measure spaces. Note in particular that, if the two "essential steps" above should turn out to be true, then we will have $\mathcal{F} = \mathcal{F}'$, and so nothing will have been lost. What is unpleasant, of course, is that one does not know what, if anything, this last step is supposed to be protecting one from.

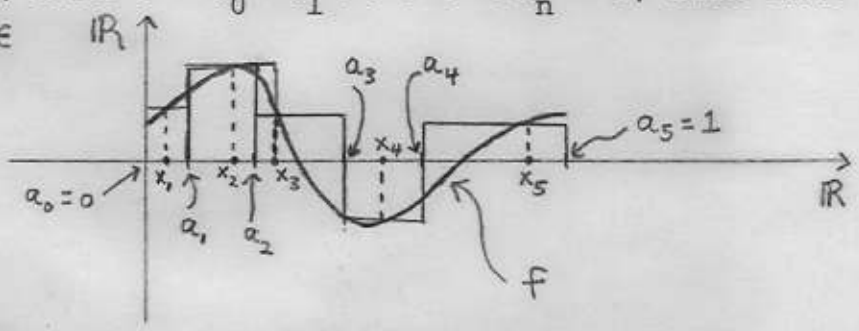
Example. Consider the product of Lebesque measure space with itself. Then the set X is the plane, $\mathbb{R}^2$. Measurable subsets include countable unions of rectangles, their complements, countable intersections of such, etc. Thus, in particular, the disk of radius r is measurable. It's measure is πr^2. Question: Why are the areas of plane geometrical figures described by the same sort of things (real numbers) as are lengths? Answer: Because $\mathbb{R} \otimes \mathbb{R} = \mathbb{R}$.

Example. Let $\hat{X}, \hat{\mathcal{M}}, \hat{M}, \hat{\mu}$ be Lebesque measure. Let $\hat{\hat{X}}, \hat{\hat{\mathcal{M}}}, \hat{\hat{M}}, \hat{\hat{\mu}}$ be the measure space with $\hat{\hat{X}}$ the set of positive integers, all subsets measurable, and $\hat{\hat{\mu}}(\hat{A})$ the number of elements of A. Then their product is the disjoint union of Lebesque measure with itself a countable number of times.

19. Example: Lebesque Integration

We now begin our study of integration. It is convenient, before discussing integration in its more general context, to first get a general idea of what is going on. We shall do this by first discussing in some detail a specific example: Lebesque integration. This particular example is important for at least two reasons: first, it illustrates all the essential ideas of integration in general, and, second, it is easily the most common example for applications. The Lebesque integral, in turn, is a generalization of the even more familiar Riemann integral. Thus, one can see more clearly what is involved in the notion of a Lebesque integral by comparison with that of a Riemann integral. Our plan for the next few sections, then, is this. We in the present section define Riemann and Lebesque integrals, obtain a few properties of each, and briefly compare. In the following several sections, we discuss in more detail the structure of Lebesque integration.

Fix some closed interval in the reals, say [0,1]. Let f be a real-valued function in this interval which is bounded, i.e., which satisfies the following condition: there is some number M such that $|f(x)| \leq M$ for every x in [0,1]. This function f is said to have Riemann integral α (where α is some real number) provided the following condition is satisfied: given any positive ϵ, there are numbers $a_0, a_1, \ldots, a_n$, with $0 = a_0 < a_1 < \ldots < a_n = 1$, such that $|\sum_1^n f(x_i)(a_i - a_{i-1}) - \alpha| \leq \epsilon$ whenever $x_1, \ldots, x_n$ satisfy $a_0 \leq x_1 \leq a_1$, $a_1 \leq x_2 \leq a_2$, $\ldots$, $a_{n-1} \leq x_n \leq a_n$. The familiar geometrical meaning of this definition is that illustrated in the figure. The sum $\Sigma f(x_i)(a_i - a_{i-1})$ is the sum of the areas of the rectangles shown. Thus, the definition requires that "if you tell me how close (ϵ) you want this sum to be to α, then I can tell you how finely $(a_0, \ldots, a_n)$ the interval [0,1] must be divided up in order that this will obtain".



Example. Let f be the function with action f(x) = x. Then f has Riemann integral 1/2. Let positive ϵ be given, and choose integer n with $\epsilon \geq 1/n$. Set $a_0 = 0/n$, $a_1 = 1/n$, $\ldots$, $a_n = n/n$. Then

the sum $\sum_1^n f(x_i)(a_i - a_{i-1})$, with each x_i satisfying $a_{i-1} \leq x_i \leq a_i$, cannot exceed the sum $\sum f(a_i)(a_i - a_{i-1}) = \sum_1^n (i/n)(1/n) = 1/2 + 1/2n$ (for, innthis last sum, we have chosen each x_i so that $f(x_i)$ is its maximum possible value). Similarly, our sum cannot be less than $\sum f(a_{i-1})(a_i - a_{i-1}) = \sum_1^n ((i-1)/n)(1/n) = 1/2 - 1/2n$. Thus, for any $x_1, \ldots, x_n$ with $a_{i-1} \leq x_i \leq a_1$, the number $\sum f(x_i)(a_i - a_{i-1})$ must be between ~~1/2~~/+ 1/2n and 1/2 - 1/2n. That is to say, we must have $|\sum f(x_i)(a_i - a_{i-1}) - 1/2| \leq \epsilon$.

Example. Let f be the function with action f(x) = 1 if x is rational, and f(x) = 0 if x is irrational. For no α does this f have Riemann integral α. Indeed, let $a_0, \ldots, a_n$ satisfy $0 = a_0 < a_1 < \ldots < a_n = 1$. Then, for each i, there is a rational x_i with $a_{i-1} \leq x_i \leq a_i$. For such choices of the x's, we have $\sum_1^n f(x_i)(a_i - a_{i-1}) = \sum_1^n (a_i - a_{i-1}) = 1$. However, for each i there is also an irrational x_i with $a_{i-1} \leq x_i \leq a_i$, and for these choices we have $\sum f(x_i)(a_i - a_{i-1}) = 0$. Since for any $a_0, \ldots, a_n$ we can choose the x's so that $\sum f(x_i)(a_i - a_{i-1})$ is either zero or one, there can be no number α such that always $|\sum f(x_i)(a_i - a_{i-1}) - \alpha| \leq 1/3$. Choose $\epsilon = 1/3$.

A function f (bounded, on [0,1]) is said to be Riemann integrable if f has Riemann integral α for some α. We next note the following fact. Let f have Riemann integral α, let ϵ be given, and let $a_0, \ldots, a_n$ be as in the definition above. Now let $b_0, \ldots, b_m$ be numbers satisfying $0 = b_0 < b_1 < \ldots < b_m = 1$, such that each of the a's is one of the b's (i.e., such that the b's define a finer division of [0,1]). Then necessarily $|\sum_1^n f(x_i)(b_i - b_{i-1}) - \alpha| \leq \epsilon$ whenever $b_{i-1} \leq x_i \leq b_i$. It follows from this observation that the Riemann integral is unique (if it exists). Indeed, let f have Riemann integral α and Riemann integral α'. Fix any positive ϵ. Let $a_0, \ldots, a_n$ and $a_0', \ldots, a_{n'}'$ be as in the definition (for α and α', respectively). Let $b_0, b_1, \ldots, b_m$ consist of all the a's and a' 's, ordered by their size. Then, whenever $b_{i-1} \leq x_i \leq b_i$, we have (by the observation above) both $|\sum f(x_i)(b_i - b_{i-1}) - \alpha| \leq \epsilon$ and $|\sum f(x_i)(b_i - b_{i-1}) - \alpha'| \leq \epsilon$. Hence, $|\alpha - \alpha'| \leq 2\epsilon$. Since this is true for every positive ϵ, we have $\alpha = \alpha'$. If f has Riemann integral α (unique, as just shown), we often write $\alpha = \int f$.

Why have we restricted consideration to functions which are bounded, and to functions defined only on some closed interval? Were

these restrictions relaxed, one would have essentially three ways by which a function could fail to be Riemann integrable: 1. The function, although defined on a closed interval, could "blow up too fast", e.g., the function f on [0,1] with $f(x) = 1/x$ (for $x \neq 0$, and with $f(0) = 0$). 2. The function, although bounded, could "fall off asymptotically too slowly", e.g., the function f on $\mathbb{R}$ with $f(x) = (1 + x^2)^{-1/2}$. (For large x, f(x) goes like $1/x$.) 3. The function could "jump around too much to have a Riemann integral", although bounded, and defined on a closed interval, e.g., the function in the example on the previous page. The first two methods of failure are not of immediate interest here: they will be taken care of more neatly later. We are here interested only in the possibilities for "jumping around", and so we consider only bounded functions on a closed interval.

Which functions are Riemann integrable, and which are not? It turns out that there is a remarkably simple criterion. Let f be a bounded function on [0,1], and let x be a number in [0,1]. For each positive ϵ, consider the number $\sup_{\epsilon} |f(x') - f(x'')|$, where the supremum is over all x' and x'' with $|x - x'| \leq \epsilon$ and $|x - x''| \leq \epsilon$. [Since f is bounded by M, this number is less than or equal to 2M.] This number represents the "total spread in f-values within ϵ of x". Clearly, this number is non-increasing as ϵ decreases. Therefore, $\lim_{\epsilon \to 0} [\sup_{\epsilon} |f(x') - f(x'')|]$ exists (and is some number between zero and 2M). We write this number $\hat{f}(x)$, thus defining a real-valued function $\hat{f}$ on [0,1]. Intuitively, $\hat{f}(x)$ is the "total spread in f-values arbitrarily close to x". [For example, if $f(x) = x$, then $\hat{f}(x) = 0$. If f has action $f(x) = 0$ for $0 \leq x \leq 1/2$, and $f(x) = 1$ for $1/2 < x \leq 1$, then $\hat{f}(x) = 0$ for $x \neq 1/2$, and $\hat{f}(1/2) = 1$, the "jump in f at 1/2".] The function f is said to be <u>continuous</u> at x if $\hat{f}(x) = 0$. [One notices that this is just a fancier way of saying the usual definition of continuity.]

<u>Example</u>. In the first example of this section, $\hat{f}(x) = 0$ for all x, and so f is continuous at every x. In the second example, $\hat{f}(x) = 1$ for all x, and so f is continuous at no x.

We now come to our criterion for Riemann integrability.

<u>Theorem</u> 13. Let f be a bounded function on [0,1]. Then f is Riemann integrable if and only if the set of points of [0,1] at which f is not continuous has (Lebesque) measure zero.

Proof: For each positive δ, denote by L_δ the set of all x with $\hat{f}(x) \geq \delta$. Then, for $\delta \leq \delta'$, we have $L_\delta \supset L_{\delta'}$, while the union of all L 's is precisely the set of points x at which f is not continuous.

Suppose first that f has Riemann integral α. Fix positive δ. Let ϵ be given, and let $a_0, a_1, \ldots, a_n$ be as in the definition. Then, whenever $a_{i-1} \leq x_i \leq a_i$, we have $\left|\sum f(x_i)(a_i - a_{i-1}) - \alpha\right| \leq \epsilon$. Denote by I the collection of all i such that (a_{i-1}, a_i) contains some point of L_δ. Then, by varying the choice of the x's, the sum $\sum_I f(x_i)(a_i - a_{i-1})$ can change by at least $\delta[\sum_I (a_i - a_{i-1})]$. Hence, by varying the choice of x's, the sum $\sum_I f(x_i)(a_i - a_{i-1})$ can also change by at least $\delta[\sum_I (a_i - a_{i-1})]$. But, since we must always have $\left|\sum_I f(x_i)(a_i - a_{i-1}) - \alpha\right| \leq \epsilon$, the sum on the left cannot vary by more than 2ϵ. We conclude: $\delta[\sum_I (a_i - a_{i-1})] \leq 2\epsilon$, i.e., $\sum_I (a_i - a_{i-1}) \leq 2\epsilon/\delta$. Now consider the union $\bigcup_I (a_i, a_{i-1})$ of open intervals. Except possibly for a finite number of points (the a_i's themselves), L_δ is contained as a subset of this union. Hence, by definition of outer measure, $\bar{\mu}(L_\delta) \leq \sum_I (a_i - a_{i-1})$. That is to say, $\bar{\mu}(L_\delta) \leq 2\epsilon/\delta$. Since ϵ is arbitrary, $\bar{\mu}(L_\delta) = 0$, i.e., L_δ is measurable and has measure zero (i.e., is null). This is true for every δ. But $L_{1/2} \cup L_{1/4} \cup L_{1/8} \cup \ldots$ is precisely the set of points at which f is not continuous. But this set, as the union of a countable collection of null sets, is itself null, i.e., it has measure zero.

Now suppose, conversely, that f is continuous except on a set of measure zero. We shall make use of the following fact: given any subset C of [0,1] which is closed (i.e., which contains all limits of sequences in C), and given any collection of open intervals which covers C, then some finite number of these open intervals also covers C. Now fix positive ϵ and δ. Then L_δ, as a subset of the set of points at which f is not continuous, has measure zero. Hence, there is a covering of L by open intervals, the sum of whose lengths is less than or equal to ϵ. But L_δ is closed. [Proof: Let $x_1, x_2, \ldots$ in L_δ converge to x. Then every neighborhood of x contains some x_i, and hence, since each x_i is in L, f varies within that neighborhood by at least δ. Hence, $\hat{f}(x) \geq \delta$, i.e., x itself is in L_δ.] Hence, by the fact above, a finite collection of these open intervals covers L_δ. We can take them to be disjoint (since the union of two overlapping open intervals is again an open interval). We conclude: there are

disjoint open intervals $A_1, \ldots, A_n$ in [0,1] which cover L_δ, and the sumoof whose lengths is $\leq \epsilon$.

Now set $B = (A_1 \ \ldots \ A_n)^c$. Then B is closed (for any sequence in B, converging to x in [0,1], could hardly have x in one of the A_i's, for then the sequence would have had to enter that A_i. Hence, x itself must be in B.) By construction, each point x of B is not in L_δ, i.e., for any x in B, $\hat{f}(x) < \delta$. That is, for each x in B, there is some open interval containing x in which f varies by less than δ. Constructing such an open interval for each such x, we obtain a covering of B by open intervals. By our fact above, a finite number covers B. As before, we may take them to be disjoint: denote them $B_1, \ldots, B_m$.

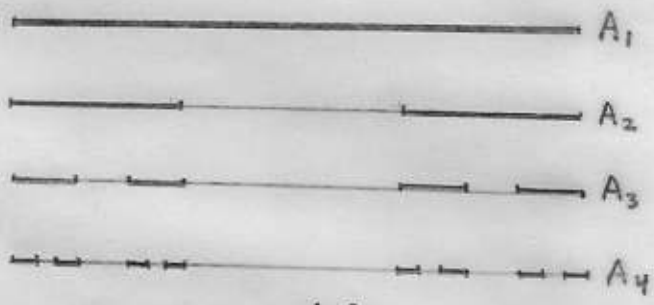
Now let $0 = a_0 < a_1 < \ldots < a_n = 1$ be the endpoints of all the A-intervals and B-intervals. For $a_{i-1} \leq x_i \leq a_i$, consider $\sum_1^n f(x_i)(a_i - a_{i-1})$. Break this sum into two parts, one over i's with (a_{i-1}, a_i) an A-interval, and the other over i's with (a_{i-1}, a_i) a B-interval. Then the first sum cannot change by more than $2M\epsilon$ as the x's change (since the total lengths of the A-intervals is $\leq \epsilon$, while, since f is bounded by M, f nowhere varies by more than 2M). The second sum cannot change by more than δ as the x's change (since the sum of the lengths of the B-intervals is certainly ≤ 1, while f varies by no more than δ on each B-interval). We conclude, therefore, that $\sum_1^n f(x_i)(a_i - a_{i-1})$ varies by no more than $2M\epsilon + \delta$ as the x's change. Since both ϵ and δ are arbitrary, we conclude: given any positive ϵ, there are $0 = a_0 < a_1 < \ldots < a_n = 1$ such that $\sum_1^n f(x_i)(a_i - a_{i-1})$ varies by no more than ϵ as the x's, with $a_{i-1} \leq x_i \leq a_i$, vary.

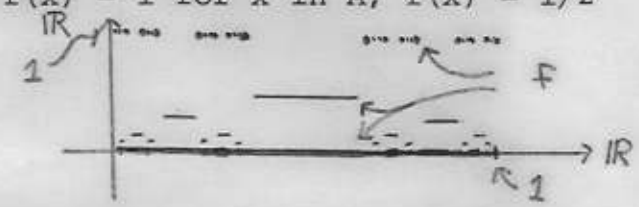
In the last statement above, set $\epsilon = 1/2$, and find $a_0, \ldots, a_n$; then set $\epsilon = 1/4$, and find $a_0', \ldots, a_n'$; then set $\epsilon = 1/8$, . . . For each of these choices of ϵ, choose any legal x's, and evaluate $\sum_1^n f(x_i)(a_i - a_{i-1})$. We thus obtain a sequence of real numbers. But by the last statement above (together with our earlier observation that finer divisions do not increase the variation in our sum), this sequence must converge to some real number α. Thus, f has Riemann integral α. /

Note, e.g., that the conclusions in our two earlier examples are immediate consequences of Theorem 13. What is striking, to me, about Theorem 13 is that its statement uses Lebesque measure, something which

does not enter directly in the definition of the Riemann integral.

Example. Let $A_1 = [0,1]$, $A_2 = A_1 \cap (1/3,2/3)^c$, $A_3 = A_2 \cap (1/9,2/9)^c \cap (7/9,8/9)^c$, etc. [That is, to obtain A_{i+1}, remove the "middle third" from each of the closed intervals which make up A_i.] Set $A = \cap A_i$. This A, called the Cantor set is measurable (since each A_i is), and in fact has measure zero (since $\mu(A_i) = (2/3)^{i-1}$). Now let f be the following function on [0,1]: f(x) = 1 for x in A; f(x) = 1/2 for x in $A_1 \cap A_2^c$, f(x) = 1/4 for x in $A_2 \cap A_3^c$, etc. This function is continuous at all x not in A. Hence, by Theorem 13, this function is Riemann integrable. [It's Riemann integral is 1/4.]





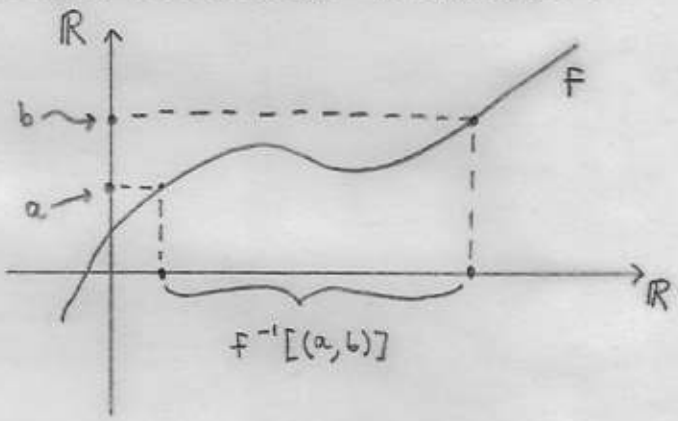
We next introduce a second kind of integral: the Lebesque integral. Fix, once and for all, a subset A of the real line, with $\mu(A)$ finite. [One could, e.g., choose A = [0,1], but we shall later need this more general case.] Let f be a bounded function on A. This function f is said to have Lebesque integral α provided the following condition is satisfied: given any positive number ϵ, there is a finite partition, $A_1, \ldots, A_n$ of A such that $\left|\sum_1^n f(x_i)\,\mu(A_i) - \alpha\right| \leq \epsilon$ whenever x_1 is in A_1, x_2 is in A_2, etc. If f has Lebesque integral α for some α, we say that f is Lebesque integrable.

The following is immediate from the fact that every closed interval is measurable: if f is Riemann integrable on [0,1], then f is also Lebesque integrable. Hence, by Theorem 13, every continuous function on [0,1] is Lebesque integrable. Note also that a function which is Riemann (and therefore also Lebesque) integrable has the same Riemann as Lebesque integral. That Lebesque integration actually "extends" Riemann integration is seen in the following.

Example. Let f be the function in the example on page 71. Then f is Lebesque integrable. Indeed, let A_1 be the rationals, and A_2 the irrationals, so $\mu(A_1) = 0$, and $\mu(A_2) = 1$. Then, for any x_1 in A_1 and

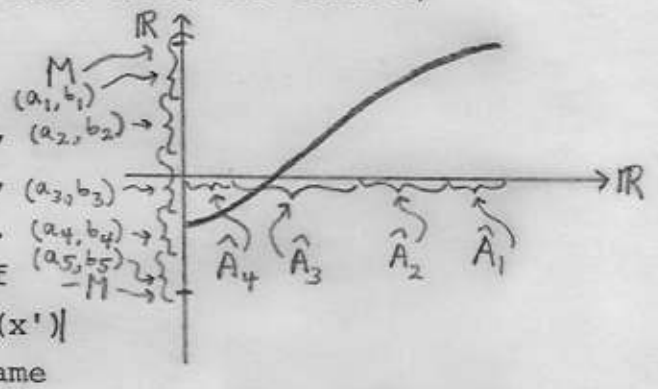
x_2 in A_2, $f(x_1)\,\mu(A_1) + f(x_2)\,\mu(A_2) = 0$. Hence, f has Lebesque integral zero.

Theorem 13 gives a simple criterion for Riemann integrability. Is there a similar criterion for Lebesque integrability? We next obtain such. We shall need the following definition. A function f defined on a measurable subset of the reals will be called measurable if, for each open interval (a,b), $f^{-1}[(a,b)]$ is a measurable set. Thus, for example, the function with action $f(x) = x$ is measurable, for, for any (a,b), $f^{-1}[(a,b)] = (a,b)$ is measurable.



Theorem 14. Let f be a bounded function defined on set A of finite measure. Then f is Lebesque integrable if and only if f is measurable.

Proof: Suppose first that f is measurable. Let positive ϵ be given. Choose open intervals (a_1,b_1), (a_2,b_2), . . . , (a_n,b_n) which cover [-M, M] (where M is the bound of f), such that, for each i, $|b_i - a_i| \leq \epsilon$. Set $\hat{A}_i = f^{-1}[(a_i,b_i)]$, so, by measurability, each of $\hat{A}_1$, . . . , $\hat{A}_n$ is measurable, and these sets cover A. Set $A_1 = \hat{A}_1$, $A_2 = \hat{A}_2 \cap \hat{A}_1^c$, $A_3 = \hat{A}_3 \cap \hat{A}_2^c \cap \hat{A}_1^c$, etc. Then A_1, . . . , A_n is a partition of A, while, by construction, $|f(x) - f(x')| \leq 2\epsilon$ whenever x and x' are in the same A_i. Now let x_i be in A_i (i = 1,2,...,n), and consider $\sum_1^n f(x_i)\,\mu(A_i)$. Since f varies on each A_i by no more than 2ϵ, this sum can vary by no more than $2\epsilon\,\mu(A)$ as the x_i change. We conclude (replacing ϵ by $2\epsilon/\mu(A)$): given $\epsilon > 0$, there is a partition A_1, . . . , A_n of A such that $\sum_1^n f(x_i)\,\mu(A_i)$ varies by no more than ϵ as the x_i (in A_i) change.



Choose $\epsilon = 1/2$, and find A_1, . . . , A_n as above; choose x_i in A_i, and set $\alpha_1 = \sum_1^n f(x_i)\,\mu(A_i)$. Choose $\epsilon = 1/4$, and find A_1', . . . , A_n', as above; choose x_i in A_i', and set $\alpha_2 = \sum_1^n f(x_i)\,\mu(A_i')$; Etc.

Then, by construction, $\alpha_1, \alpha_2, \ldots$ is a Cauchy sequence of real numbers (in fact, for $n \geq N$ and $n' \geq N$, we have $|\alpha_n - \alpha_{n'}| \leq 2^{-N+1}$). Hence, this sequence converges to some number α. But $|\alpha - \alpha_1| \leq 1$, $|\alpha - \alpha_2| \leq 1/2$, etc. Clearly, then, f has Lebesque integral α.

Now suppose, conversely, that f is Lebesque integrable, with Lebesque integral α. Let (a,b) be any open interval: we must show that $K = f^{-1}[(a,b)]$ is measurable. Fix positive ϵ, and let $A_1, \ldots, A_n$ be the partition of A such that $|\sum_1^n f(x_i)\, \mu(A_i) - \alpha| \leq \epsilon$ whenever x_i is in A_i $(i = 1,\ldots,n)$. Then, in particular, the number $\sum_1^n f(x_i)\, \mu(A_i)$ varies by no more than 2ϵ as the x's change.

Fix positive δ. Denote by I the collection of all integers i such that the value of f on A_i varies by no more than δ, and by J the collection of i such that f varies by more than δ on A_i. Then we must have $\delta(\sum_J \mu(A_i)) \leq 2\epsilon$, i.e., $\mu(\bigcup_J A_i) \leq 2\epsilon/\delta$. Set $L = \cup A_i$, where the union is over all i in I such that, for some x in A_i, f(x) is in the open interval $(a+\delta, b-\delta)$. Then, since f varies by no more than δ on each A_i with i in I, f(x) is in (a,b) for each x in L. That is to say, $L \subset K$. Set $M = \bigcup_J A_i$. Then, as we saw above, $\mu(M) \leq 2\epsilon/\delta$. Finally, note that $L \cup M \supset f^{-1}(a+\delta, b-\delta)$.

To summarize, we conclude (replacing ϵ by $\epsilon/2\delta$): given positive ϵ and δ, there are disjoint measurable sets L and M such that $L \subset (a,b)$, $f[L \cup M] \supset f^{-1}(a+\delta, b-\delta)$, and such that $\mu(M) \leq \epsilon$.

Set $\epsilon = \bar{\epsilon}$ and $\delta = 1$; then set $\epsilon = \bar{\epsilon}/2$, and $\delta = 1/2$; then set $\epsilon = \bar{\epsilon}/4$, and $\delta = 1/4$; etc. In each case, determine L and M as above. Denote by $\bar{L}$ the union of the L's so obtained, and by $\bar{N}$ the union of the $(L \cup M)$'s. Then, clearly, $\bar{L} \subset K \subset \bar{N}$, while $\mu(\bar{N} \cap \bar{L}^c) \leq 2\bar{\epsilon}$. [That is to say, we have now "sandwiched" K between measurable sets with their difference having small measure.] Now set $\bar{\epsilon} = 1$; then $\bar{\epsilon} = 1/2$; then $\bar{\epsilon} = 1/4$; etc. Denote by L' the union of the corresponding $\bar{L}$'s, and by N' the intersection of the corresponding $\bar{N}$'s. Then we have $L' \subset K \subset N'$, with $\mu(N' \cap L'^c) = 0$. But Lebesque measure is complete: hence, the set K is measurable.

Since, for every open interval (a,b), $K = f^{-1}[(a,b)]$ is measurable, the function f is measurable. /

The proofs of the two halves of Theorem 13 and of the two halves of Theorem 14 are all based on essentially the same idea.

Example. Every continuous function on [0,1] is measurable. Fancy proof: every continuous function is Riemann integrable, by Theorem 13; hence, is Lebesque integrable; hence is measurable, by Theorem 14.
Example. Let B be the non-measurable subset of [0,1] given in the example on page 24. Let f be the function on [0,1] with $f(x) = 1$ if f is in B, and $f(x) = 0$ otherwise. By Theorem 14, this f is not Lebesque integrable (since, e.g., $f^{-1}[(1/2,3/2)] = B$ is not measurable). We show this directly. Let $A_1,\ldots,A_n$ be a partition of [0,1]. Then, since $\overline{\mu}(B) = 1$, each A_i either contains a point of B, or else has measure zero; since $\overline{\mu}(B^c) = 1$, each A_i either contains a point of B^c, or else has measure zero. Hence, for x_i in A_i, the number $\sum_i^n f(x_i)\,\mu(A_i)$ can assume either of the values zero or one. Hence, e.g., it is impossible that there be a number α such that always $\left|\sum_i^n f(x_i)\,\mu(A_i) - \alpha\right| \leq 1/3$.

It should now be clear what is going on here. The Riemann integral requires a partition by intervals, the Lebesque integral by any measurable sets. As a consequence, more functions are Lebesque than Riemann integrable. [Both Riemann and Lebesque wish to find the total value of a pile of coins on a table. Riemann would find the value of the first ten coins he picks up, then of the next ten, etc., and would then add the results. Lebesque would first find the total value of the pennies, then the total value of the nickles, etc., and would add.]

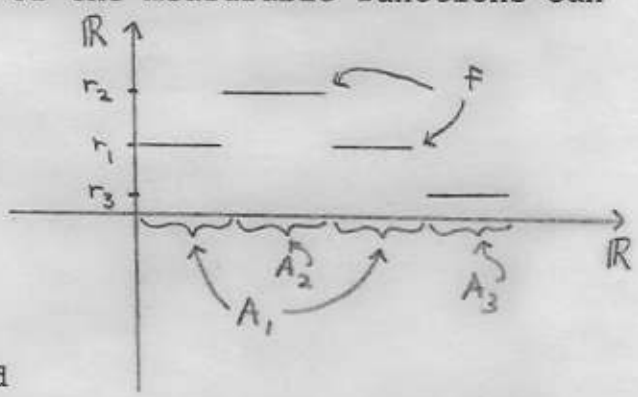
It might appear that the only value of the Lebesque over the Riemann integral is that one can thereby integrate such uninteresting functions as that on page 71. It will turn out, however, that the Lebesque integral enjoys a much more important advantage: Lebesque integrable functions remain Lebesque integrable under, e.g., limits, whereas Riemann integrable functions do not. That is to say, one can more freely do operations on functions, preserving integrability, in the Lebesque than the Riemann case. The essential difference between the two will be seen to ultimately arise from the following: the Lebesque measurable subsets of the reals form a measurable space, while the intervals, as subsets of the reals, do not form a measurable space.

20. Measurable Functions; Step Functions (Lebesque)

We saw in the previous section (Theorem 14) that measurability is the criterion for Lebesque integrability of a bounded function defined on a set of finite measure. This criterion may not, at first glance, appear to be very useful. How does one determine whether or not a function is measurable? One must consider all open intervals, (a,b), and, for each, decide whether or not $f^{-1}[(a,b)]$ is measurable: a possibly formidable job. What one might like to do, then, is i) develop some properties of measurable functions, and ii) develop an intuitive feeling for which functions are measurable. This is our purpose in the present section.

Let f be a measurable function. For any sets $A_1, A_2, \ldots$, we have $f^{-1}[\cup A_i] = \cup f^{-1}[A_i]$, $f^{-1}[\cap A_i] = \cap f^{-1}[A_i]$, and $f^{-1}[A_i^c] = (f^{-1}[A_i])^c$. Letting $A_1, A_2, \ldots$ be open intervals, and recalling that complements, together with countable unions and intersections, of measurable sets are measurable, we have: for f measurable, and A any set which can be obtained by the operations of complements, countable unions, and countable intersections from open intervals, $f^{-1}[A]$ is measurable. Thus, for example, $f^{-1}[[a,b]]$, $f^{-1}[(a,b]]$, $f^{-1}[a]$, $f^{-1}[(-\infty,b)]$, and $f^{-1}[(-\infty,b]]$ are all measurable sets. Note also that, for measurability of f, it suffices to check that each $f^{-1}[(a,b)]$ is measurable for, e.g., a and b rational.

A more useful characterization of the measurable functions can be obtained as follows. A step function is a measurable function with finite range, i.e., a measurable function such that f(x) assumes only a finite number of different values. Clearly, every step function (on measurable set A) can be constructed as follows: choose a finite partition $A_1, \ldots, A_n$ of A, and n numbers $r_1, \ldots, r_n$. Let f have action $f(x) = r_i$ for x in A_i. The range of this step function f is then the set consisting of $r_1, \ldots, r_n$.



Measurable functions can be described in terms of step functions as follows:

Theorem 15. Let f be a bounded function defined on set A of finite

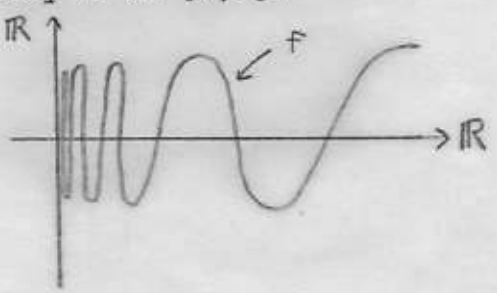
measure. Then f is measurable if and only if f can be approximated by a step function in the following sense: given any positive ϵ, there is a step function $\hat{f}$ such that $|f(x) - \hat{f}(x)| \leq \epsilon$ for every x in A.

Proof: Suppose first that f is measurable. Let positive ϵ be given. Choose numbers $a_0, \ldots, a_n$ with $-M = a_0 < a_1 < \ldots < a_n = M$, where M is the bound of f, and with $|a_{i+1} - a_i| \leq \epsilon$ for each i. For each i, set $A_i = f^{-1}[[a_{i-1}, a_i)]$ (except $A_n = f^{-1}[[a_{n-1}, a_n]]$). Then $A_1, \ldots A_n$ is a partition of A. On each A_i, f varies by no more than ϵ. Fix a point x_i of each A_i. Now let $\hat{f}$ be the step function with the following action: $\hat{f}(x) = f(x_i)$ whenever x is in A_i. Clearly, then, $|f(x) - \hat{f}(x)| \leq \epsilon$ whenever x is in A.

Now suppose, conversely, that f can be approximated by step functions as in the theorem. Fix positive ϵ, and let $\hat{f}$ be the step function ϵ-approximating f. Then there is a partition $A_1, \ldots, A_n$ of A, and numbers $r_1, \ldots, r_n$ such that $\hat{f}(x) = r_i$ for x in A_i. Now consider, for any x_i in A_i, the sum $\sum_1^n f(x_i)\, \mu(A_i)$. Clearly, this sum differs by no more than $\epsilon\, \mu(A)$ from $\sum_1^n r_i\, \mu(A_i)$. That is to say, the function f is Lebesque integrable. By Theorem 14, f is measurable. /

It is obvious, even without Theorems 14 and 15, that the step functions are Lebesque integrable (and, in fact, that the integral of such a function is $\sum_1^n r_i\, \mu(A_i)$). But, by Theorems 14 and 15, the Lebesque integrable functions are precisely those which can be approximated (in the sense of Theorem 15) by step functions. In a sense, then, "once you know how to integrate step functions, you know how to integrate everything which can be integrated".

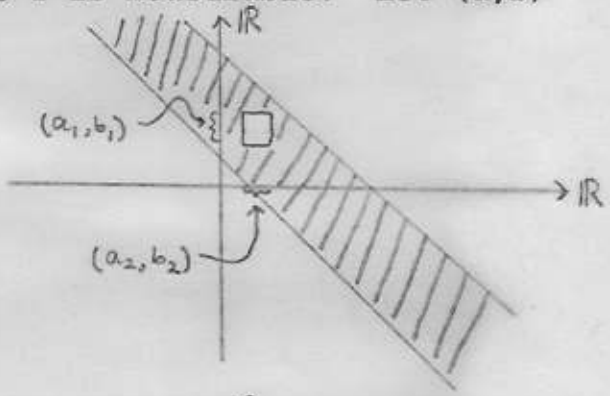
One might imagine that something analogous to Theorem 15 might also be true for Riemann integration. For example: a bounded function f on [0,1] is Riemann integrable if and only if there is a partition of [0,1] by intervals, and a function $\hat{f}$, constant on each of these intervals, such that $|f(x) - \hat{f}(x)| \leq \epsilon$ for every x in [0,1]. Unfortunately,



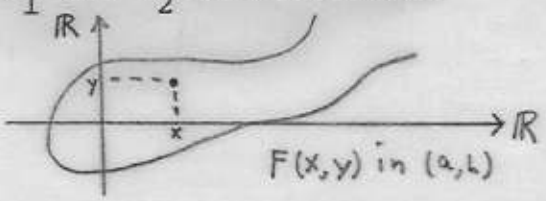
Example. Let f be the function with action $f(x) = \sin(1/x)$ for $x \neq 0$, and $f(0) = 0$. Then, since f is continuous except at x = 0, f is

Riemann integrable. Given any partition $A_1, \ldots, A_n$ of [0,1] by intervals, consider that interval, say A_n, which includes the number one. Then, on that A_n, $f(x) = 1$ for some x in A_n, while $f(x) = -1$ for other x in A_n. Hence, for no step function $\hat{f}$ based on $A_1, \ldots, A_n$ can we have $|f(x) - \hat{f}(x)| \leq 1/2$ for every x in A.

Which operations yield measurable functions when applied to measurable functions? Let f be measurable, let a be any real number, and consider the function af, with action $(af)(x) = a(f(x))$. Then af is measurable, for, given open interval (b,c), $(af)^{-1}[(b,c)] = f^{-1}[(b/a,c/a)]$. As a second example, consider sums. Let f_1 and f_2 be measurable, and let f be $f_1 + f_2$ the function with action $f(x) = f_1(x) + f_2(x)$. We show that this f is measurable. Let (a,b) be an open interval, Then $f^{-1}[(a,b)]$ consists of all x with $f_1(x) + f_2(x)$ in (a,b), i.e., of all x with the point $(f_1(x), f_2(x))$ of $\mathbb{R}^2$ in the region indicated in the figure. Let (a_1,b_1) and (a_2,b_2) be open intervals such that the rectangle $(a_1,b_1) \times (a_2,b_2)$ is a subset of this region as shown. Then the point $(f_1(x), f_2(x))$ of $\mathbb{R}^2$ is in this rectangle when and only when x is in both $f_1^{-1}[(a_1,b_1)]$ and $f_2^{-1}[(a_2,b_2)]$, i.e., when and only when x is in the measurable set $f_1^{-1}[(a_1,b_1)]$ $f_2^{-1}[(a_2,b_2)]$. Since the union of a countable collection of such rectangles is precisely the region in the figure, we conclude that $f^{-1}[(a,b)]$ is the union of a countable collection of measurable sets. That is, $f^{-1}[(a,b)]$ is measurable, and hence f is measurable. [For the case when f_1 and f_2 are bounded, and defined on a set of finite measure, we could obtain the same result directly from Theorem 15, using the fact that the sum of two step functions is a step function.]



The examples above are special cases of a more general result. Let F(be a continuous function of two real variables (or, more generally, of any number of real variables). That is, let F be such that, whenever $x_1, \ldots$ converges to x and $y_1, \ldots$ converges to y, the numbers $F(x_i,y_i)$ converge to F(x,y). Let f_1 and f_2 be measurable functions, and let f be the function with action $f(x) = F(f_1(x),f_2(x))$. Then f is also measurable. The proof



is essentially the same as that for the sum above. Let K denote the subset of $\mathbb{R}^2$ consisting of all (x_1, x_2) with $F(x_1,x_2)$ in (a,b). By continuity, K is the union of a countable collection of rectangles of the form $(a_1,b_1) \times (a_2,b_2)$. Then $f^{-1}[(a,b)]$ is the union of a countable collection of measurable sets of the form $f_1^{-1}[(a_1,b_1)]$ $f_2^{-1}[(a_2,b_2)]$. Hence, $f^{-1}[(a,b)]$ is measurable, whence the function f is measurable.

By the previous result, we have, for example, the following. If f is measurable, then so is $|f|$ (the function with action $(|f|)(x) = |f(x)|$). [Proof: The function F of one variable with action $F(y) = |y|$ is continuous.] If f and g are measurable, then so is fg (the function with action $(fg)(x) = (f(x))(g(x))$). [Proof: The function F of two variables with action $F(x,y) = xy$ is continuous.]

In short, essentially every "algebraic operation" one can think of, when applied to measurable functions, yields measurable functions. [This is perhaps what one might have expected, in light of the difficulty of even finding a set which is not measurable.]

Are limits of measurable functions measurable? There are, of course, many ways in which one function can be a "limit" of a sequence of functions. As an example, we here consider one such. Let f be a function, and let $f_1, f_2, \ldots$ be a sequence of measurable function such that $\lim_{i \to \infty} f_i(x) = f(x)$ for each x. Then, we claim, the function f is also measurable. Proof: Let (a,b) be an open interval, and set $A_i = f_i^{-1}[(a,b)]$, a measurable set. Then x is in $f^{-1}[(a,b)]$ if and only if f(x) is in (a,b), i.e., if and only if, for some N, $f_n(x)$ is in (a,b) whenever $n \geq N$, i.e., if and only if, for some N, x is in A_n whenever $n \geq N$, i.e., if and only if, for some N, x is in $\bigcap_{n \geq N} A_n$, i.e., if and only if x is in $\bigcup_N \bigcap_{n \geq N} A_n$. Thus, $f^{-1}[(a,b)] = \bigcup_N \bigcap_{n \geq N} A_n$, and so f is measurable.

The following example shows that such point-wise limits of Riemann integrable functions need not be Riemann integrable.

Example. Let $r_1, r_2, \ldots$ be the rationals in [0,1]. For each n, let f_n be the function with the following action: $f_n(r_i) = 1$ if $i < n$, and $f_n(x) = 0$ otherwise. Let f be the function with $f(x) = 1$ if x is rational, and zero otherwise. Then, for each x in [0,1], $\lim f_n(x) = f(x)$. Each f_n is Riemann integrable (since it is continuous except at a finite number of points). But f is not Riemann integrable.

21. Dependence on Region of Integration (Lebesque)

There are two things on which the value of the Lebesque integral depends: the region of integration, and the function integrated. We consider these two dependences, respectively, in this and the following section.

Let $X, \mathcal{M}, M, \mu$ be Lebesque measure. Fix, once and for all, a measurable, real-valued function f on X. Then f is measurable on every measurable subset of X. Denote by $\mathcal{F}$ the collection of all measurable subsets A of X with finite measure, and on which f is bounded. Then, for A in $\mathcal{F}$, f is Lebesque integrable on A, by Theorem 14. We write the value of this integral as $\nu(A) = \int_A f$. [Use of the same symbol for the Lebesque and Riemann integral causes no confusion because i) the two are equal whenever both are defined, and ii) people who know the Lebesque integral always mean by "$\int$" the Lebesque integral.]

We are interested in the dependence of the number $\nu(A)$ on the set A. Essentially the complete list of properties is contained in the following assertion - $X, \mathcal{M}, M, \mathcal{F}, \nu$ is a pre-measure space - which we now prove. Proof: 1. ϕ is in $\mathcal{F}$, and $\nu(\phi) = 0$. 2. Let A be in $\mathcal{F}$ (so A has finite measure, and f is bounded on A), and let B be a measurable subset of A. Then certainly B has finite measure, and certainly f is bounded on B. Hence, B is in $\mathcal{F}$. 3. If A and B are in $\mathcal{F}$, then $A \cup B$ has finite measure, and f is bounded on $A \cup B$ (by the maximum of the bounds on A and B), whence $A \cup B$ is in $\mathcal{F}$. 4. Let A be in $\mathcal{F}$, and let $A_1, A_2, \ldots$ be a partition of A. Fix positive ϵ, and let g be a step function on A with $|f(x) - g(x)| \leq \epsilon$ for every x in A. Then, evidently, $\left|\int_A f - \int_A g\right| \leq \epsilon \mu(A)$, while, for every measurable subset B of A, $\left|\int_B f - \int_B g\right| \leq \epsilon \mu(B) \leq \epsilon \mu(A)$. Since ϵ is arbitrary, and since $\sum \int_{A_i} g = \int_A g$, we have: given any positive ϵ, there is a finite collection K of integers such that, for any finite K' containing K as a subset, $\left|\sum_{K'} \int_{A_i} f - \int_A f\right| \leq \epsilon$. But this is precisely the statement that $\sum \nu(A_i) = \nu(A)$.

Thus, we obtain a pre-measure space, and so, by Theorem 10, a measure space. We write this measure, so constructed, as μ_f (to show dependence on the function f). For A a measurable set, instead of $\mu_f(A)$, we sometimes write $\int_A f$ (a harmless convention since, for A in $\mathcal{F}$, $\mu_f(A) = \nu(A)$). Note that it is possible that $\mu_f(A) = \infty$ for some measurable A's.

Example. Let f be the function with action $f(x) = 1$ for all x. Then $\mu_f = \mu$.

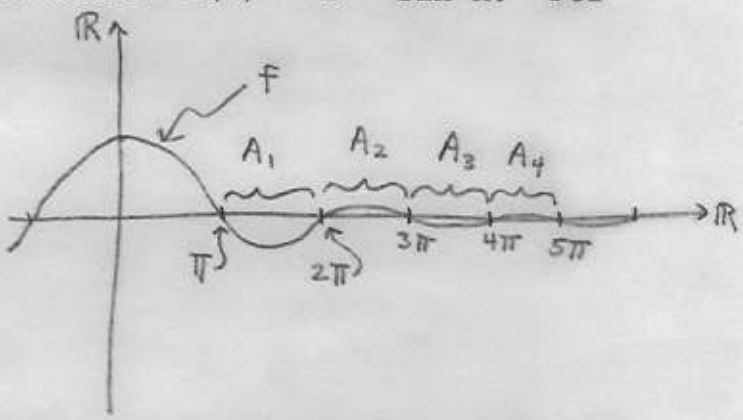
Example. Let f be the function with action $f(x) = (1 + x^2)^{-1}$. Then, e.g., $\mu_f(X) = \pi$.

Example. Let f be the function with action $f(x) = x^{-1/2} e^{-x}$. Then, e.g., $\mu_f(X) = 2\sqrt{\pi}$.

Thus, we can now integrate functions even if they are not bounded, and even if they are not defined only on a set of finite measure (as the last two examples above illustrate). The construction which leads from a pre-measure space to a measure space automatically takes care of all possible difficulties, i.e., one obtains infinite integral when that integral should "diverge", it gives the right answer for "converging" integrals, etc. Furthermore, since "dependence on the region of integration" is now that of a measure space, all the facts concerning measure spaces immediately become facts about integrals. We give one example. Let f be a measurable function on X, let $A_1 \supset A_2 \supset \ldots$ be measurable sets with empty intersection, and with $\int_{A_1} f$ finite. Then $\lim \int_{A_i} f = 0$. Proof: μ_f is a measure; use Theorem 9.

Some analysis texts make a distinction between integrals which "converge absolutely" and those which "converge conditionally". How would this distinction be expressed in the present language?

Example. Let f be the function with action $f(x) = x^{-1} \sin x$. For $i = 1,2,\ldots$, denote by A_i the interval $[i\pi, (i+1)\pi)$. Then, for i even, $(\sin x)/(i+1)\pi \leq f(x) \leq (\sin(x)/i\pi$ for every x in A_i. Hence, for i even, $2/(i+1)\pi \leq \int_{A_i} f \leq 2/i\pi$. Thus, $\sum_{\text{even } i} \int_{A_i} f \geq \sum_{\text{even}} 2/(i+1)\pi = \infty$, and therefore $\int_X f = \infty$.



For each positive number a set, in the example above, $B_a = [-a, a]$. Then, for each a, $\int_{B_a} f$ is finite (since each B_a has finite measure, and since f is bounded). Furthermore, $\lim_{a \to \infty} \int_{B_a} f$ is also finite. [Proof: By the same argument as in the example, we have that, for odd i, $-2/i\pi \leq \int_{A_i} f \leq -2/(i+1)\pi$. The result now follows from the

sequence $c_1, c_2, \ldots$ whose n^{th} term is $c_n = 1/2 - 1/3 + 1/4 - \ldots \pm 1/n$ converges.] When $\lim \int_{B_a} f$ exists and is finite, one says that f is "conditionally integrable", with integral this limit. Thus, integrable in our sense corresponds to "absolutely integrable". The other notion does not appear because it makes reference to a special structure of Lebesque measure: the fact that the underlying set X is in that case ordered. That is to say, there is no notion of "conditionally integrable" in a general measure space.

Finally, one might ask whether or not our dealing first with bounded functions on a set of finite measure, and only from this with the general case, leads to a too restrictive notion of an integral. Could there not be some measurable function f which "obviously should have $\int_X f$ finite", but which in fact has $\int_X f = \infty$ just because one is not able to find enough sets of finite measure on which f is bounded? We have, presumably, "given $\int_X f$ every reasonable opportunity to be finite" if there exists a partition of X by sets of finite measure on each of which f is bounded (for then $\int_X f = \infty$ means that it is a sum of a non-converging series with finite entries). Thus, what one would like to show is this: for every measurable function f, μ_f is sigma-finite. We claim that this is in fact true. Proof: For $i = 1, 2, \ldots$, set $A_i = f^{-1}[(-i, i)]$. Then each A_i is measurable, f is bounded (with, in fact, $M = i$) on each A_i, and the A_i cover X. But Lebesque measure is sigma-finite, and so each A_i can be written as the union of a countable collection of measurable sets with finite Lebesque measure. For each of these sets, μ_f is finite. Thus, X can be written as a countable union of sets on each of which μ_f is finite, whence the measure μ_f is sigma-finite.

22. Dependence on the Function (Lebesgue)

Let $X, \mathcal{M}, M, \mu$ be Lebesque measure. We saw in the previous section that, with each measurable function f on X, there is associated a certain measure μ_f on $X, \mathcal{M}$. We now wish to consider the dependence of the measure μ_f on the function f.

When does $\mu_f = \mu_g$? It is convenient to introduce the following definition: a statement about points of X will be said to be true almost everywhere if it is true except on some subset of X which is null (i.e., which, in our Lebesque case, has measure zero). We now claim that $\mu_{f+g} = \mu_g$ if and only if $f = g$ almost everywhere (i.e., if and only if there is some set A with measure zero such that $f(x) = g(x)$ whenever x is not in A). Proof: Suppose first that $\mu_f = \mu_g$. Let A be any set of finite measure, on which both f and g are bounded. For ϵ positive, denote by A_ϵ the subset of A consisting of points x at which $f(x) \geq g(x) + \epsilon$. Then $\mu_f(A) \geq \mu_g(A) + \epsilon\mu(A_\epsilon)$, whence, since $\mu_f = \mu_g$, we must have $\mu(A_\epsilon) = 0$. Hence, $A_1 \cup A_{1/2} \cup A_{1/4} \cup \ldots$ has measure zero. That is, $f \geq g$ on A almost everywhere. Since similarly with "$f \leq g$", we have: $f = g$ almost everywhere on A. But, since X is a countable union of A's of finite measure on which f and g are bounded, we have $f = g$ almost everywhere on X. Suppose next that $f = g$ almost everywhere on X. By sigma-finiteness, it suffices, in order to show $\mu_{f+g} = \mu_g$, to show that $\mu_f(A) = \mu_g(A)$ for any A of finite measure on which both f and g are bounded. But this is obvious, for, given any partition $A_1, \ldots, A_n$ of A, there is a finer partition $A'_1, \ldots, A'_{n+1}$ such that $f = g$ on $A'_1 \ldots A'_n$, and such that $\mu(A'_{n+1}) = 0$.

We next note the following two obvious facts. If measurable f is bounded by M on set A of finite measure, then $\left|\int_A f\right| \leq M\mu(A)$. If f is measurable, and a is any real number, then $\mu_{af} = a\mu_f$.

How do integrals behave under sums of functions?

Example. Let f be the function with action $f(x) = 1$ for all x, and g the function with $g(x) = -1$. Then $\mu_f = \mu$ and $\mu_g = -\mu$. Hence, $\mu_f(X) = \infty$ and $\mu_g(X) = \infty$, whence $\mu_f(X) + \mu_g(X) = \infty$. On the other hand, $f + g$ is the zero function, whence $\mu_{f+g}(X) = 0$.

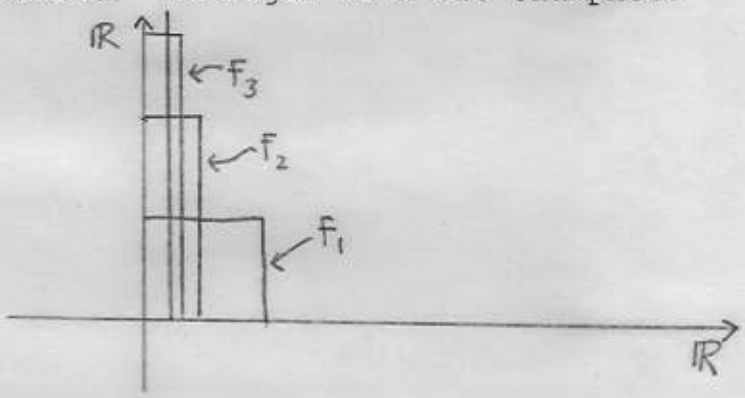
Thus, not suprisingly, one cannot add under the integral when divergent integrals are involved. The characterization of behavior under addition of functions, as suggested by the above example, is this: for f and

g measurable, $\mu_{f+g} = \mu_f + \mu_g$ [Recall, from Section 15, the definition of the sum of two measures.] Proof: It suffices to show that $\mu_{f+g}(A) = \mu_f(A) + \mu_g(A)$ for any set A of finite measure on which both f and g are bounded. Let A be such a set. Then, clearly, $\mu_{f'+g'}(A) = \mu_{f'}(A) + \mu_{g'}(A)$ when f' and g' are step functions. But, given positive ϵ, we can find such step functions with $|f' - f| \leq \epsilon$ and $|g' - g| \leq \epsilon$. This having been done, we have $|\mu_{f'}(A) - \mu_f(A)| \leq \epsilon\mu(A)$, $|\mu_{g'}(A) - \mu_g(A)| \leq \epsilon(A)$, and $|\mu_{f'+g'}(A) - \mu_{f+g}(A)| \leq 2\epsilon\mu(A)$. Since ϵ is arbitrary, the result follows.

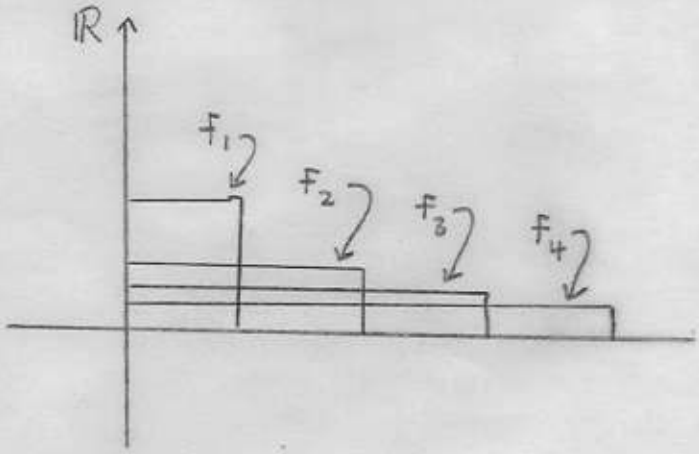
Thus, under addition of functions, the corresponding measures add. Note that we originally defined the sum of two measures, not to make such a result as this be true, but rather in the only reasonable way so that the sum would again be a measure. It is comforting that sums of measures turns out to be also the appropriate notion for addition under integral signs.

The final issue is that of limits. We begin with two examples.

Example. For i = 1, 2, . . . let f_i be the function with action $f_i(x) = i$ for x in the open interval (0, 1/i), and $f_i(x) = 0$ otherwise. Then, for each x, $\lim f_i(x) = f(x)$, where f is the zero function. Let A = [0,1]. Then $\int_A f_i = 1$ for each i, while $\int_A f = 0$.



Example. For i = 1, 2, . . . let f_i be the function with action $f_i(x) = 1/i$ for x in the open interval (0, i), and $f_i(x) = 0$ otherwise. Then, for each x, $\lim f_i(x) = f(x)$, where f is the zero function. We have $\int_X f_i = 1$ for each i, while $\int_X f = 0$.



In the first example, each f_i is bounded, the set A over which the integrals are performed has finite measure, but there is no single

bound which works for all the f_i simultaneously. In the second example, the f_i have a common bound ($M = 1$), but the integrals are over a set (X) of infinite measure. One might imagine, therefore, that one can take limits under the integral sign whenever the functions have a common bound, and the integral is over a set of finite measure. Not only is this true, but in fact there is the following much stronger result.

Theorem. 16. Let φ and $f_1, f_2, \ldots$ be measurable functions such that, for every i and x, $|f_i(x)| \leq \varphi(x)$, and such that $\int_X \varphi$ is finite. Let f be a function with $f(x) = \lim f_i(x)$ for every x. Then $\int_X f = \lim \int_X f_i$.

Proof: First note that $|f| \leq \varphi$. For each positive integer n, let ψ_n be the function with $\psi_n(x) = \sup_{i \geq n} |f_i(x) - f(x)|$. Then each ψ_n is non-negative and (since $|\psi_n| \leq 2\varphi$) is integrable over X. We have, for each x, $\psi_1(x) \geq \psi_2(x) \geq \ldots$, and $\lim \psi_n(x) = 0$. Let A be a set with finite measure, and on which ψ_1 (and hence each ψ_n) is bounded, say by M. Fix positive ϵ, and let, for each n, A_n be the subset of A consisting of x with $\psi_n(x) \geq \epsilon$. Then $A_1 \supset A_2 \supset \ldots$, and $\cap A_n = \emptyset$. Hence, by Theorem 9, $\mu(A_N) \leq \epsilon$ for some N. Then, for $n \geq N$, we have $\int_A \psi_n \leq \epsilon M + \epsilon \mu(A)$. Since ϵ is arbitrary, we conclude: $\lim \int_A \psi_n = 0$. Since X can be covered by a countable collection of such A's, we have $\lim \int_X \psi_n = 0$. But, for each n, $\psi_n \geq |f_n - f|$, and so $\lim \int_X |f_n - f| = 0$. But $\left|\int_X f_n - \int_X f\right| \leq \int_X |f_n - f|$, and so $\lim \left|\int_X f_n - \int_X f\right| = 0$. That is to say, $\lim \int_X f_n = \int_X f$. /

Consider now the case when each f_i is defined on some set A of finite measure, and the f_i have a common bound M. Let φ be the function on A with $\varphi(x) = M$ for every x in A. This φ then satisfies the conditions of Theorem 16, and so we can again take limits under integral signs.

Theorem 16 is by far the most useful statement about when the limit of the integral is the integral of the limit.

23. The Space of Measurable Functions (Lebesque)

In the last three sections, we have introduced the Lebesque integral, together with some of its properties. Our purpose in the present section is to reformulate this work in a way more elegant, more dramatic, and more memorable.

Let $X, \mathcal{M}, M, \mu$ be Lebesque measure. Denote by $\mathcal{F}$ the collection of all real-valued measurable functions on $X, \mathcal{M}$, where two functions are taken as defining the same element of $\mathcal{F}$ if they are equal almost everywhere. [That is, $\mathcal{F}$ is the set of equivalence classes of measurable functions under the equivalence relation "equal almost everywhere".] Define addition in $\mathcal{F}$ by adding the corresponding functions pointwise (noting first that the sum of two measurable functions is measurable, and second that this sum in $\mathcal{F}$ is well-defined: if $f = f'$ almost everywhere and $g = g'$ almost everywhere, then $f + g = f' + g'$ almost everywhere). Clearly, $\mathcal{F}$ is an abelian group under this addition. [A representative of zero in $\mathcal{F}$ is the zero function. The inverse of the element of $\mathcal{F}$ represented by f is the element represented by $-f$.]

Next, consider $SF(X, \mathcal{M}, M)$, the complete abelian topological group of all sigma-finite measures on $X, \mathcal{M}$, valued in M. We denote by μ_* the following mapping from $\mathcal{F}$ to $SF(X, \mathcal{M}, M)$: for f a representative of an element of $\mathcal{F}$, $\mu_*(f)$ is the sigma-finite measure μ_f. [Note that this is well-defined: if $f = g$ almost everywhere, then $\mu_f = \mu_g$.]

What properties has this mapping μ_*? First note that it is one-to-one: if $\mu_f = \mu_g$, then $f = g$ almost everywhere. We also have from Section 22 that $\mu_{f+g} = \mu_f + \mu_g$ and $\mu_{-f} = -\mu_f$. That is to say, the mapping μ_* from $\mathcal{F}$ to $SF(X, \mathcal{M}, M)$ satisfies the first condition (page 40) for a homomorphism.

Does this μ_* satisfy the second condition for a homomorphism? We cannot answer this question immediately, for $\mathcal{F}$ is so far only an abelian group, rather than an abelian topological group, while the 0-neighborhoods in $\mathcal{F}$ are necessary to give meaning to the second condition. What one would like to do, therefore, is to introduce some 0-neighborhoods in $\mathcal{F}$, and to do so in such a way that μ_* is indeed a homomorphism. As an example of such an attempt, we try the following: a subset U of $\mathcal{F}$ is a 0-neighborhood if there exists

a positive number ϵ such that every element of $\mathcal{F}$ having representative f with $|f| \leq \epsilon$ almost everywhere is in U. It is easily checked that, with these 0-neighborhoods, $\mathcal{F}$ is indeed an abelian topological group. With these 0-neighborhoods for $\mathcal{F}$, is $\mu_\cdot$ a homomorphism? Unfortunately, the answer is no. Let positive ϵ be given, and denote by V the corresponding 0-neighborhood in $SF(X, \mathcal{M}, M)$. That is, $\tilde{\mu}$ is in V provided $|\tilde{\mu}(A)| \leq \epsilon$ for every measurable A. Then $\mu_\cdot^{-1}[V]$ consists of all measurable f such that μ_f is in V, i.e., of all measurable f such that $|\mu_f(A)| \leq \epsilon$ for every measurable A. Does there exist a positive number δ such that, whenever $|f| \leq \delta$ almost everywhere, $|\mu_f(A)| \leq \epsilon$ for every measurable A? The answer is no, for, choosing for f the constant function with value δ, $|\mu_f(A)|$ can be as large as we please for suitable A. Thus, the second condition for a homomorphism fails.

The difficulty above was that we made the wrong choice for the 0-neighborhoods in $\mathcal{F}$. The "right" choice is the following. A subset U of $\mathcal{F}$ is a 0-neighborhood if there exists positive ϵ such that every element of $\mathcal{F}$ having representative f with $\int_X |f| \leq \epsilon$ is in U. [Intuitively, the first choice of 0-neighborhoods looks only at how large the values of f are, while the present choice cares also how long f cares to maintain a large value.] We now claim that, with these 0-neighborhoods, $\mathcal{F}$ is an abelian topological group. Proof: 1. If $|\int_X |f|| = 0$, then f = 0 almost everywhere. 2,3,4. Obvious. 5. Let U be a 0-neighborhood of $\mathcal{F}$, and let ϵ be a corresponding positive number. Let V consist of all f with $\int_X |f| \leq \epsilon/2$. Then, for f and f' in V, $\int_X |f + f'| \leq \int_X |f| + |f'| = \int_X |f| + \int_X |f'| \leq \epsilon$. Thus, $V + V \subset U$.

We now have an abelian topological group $\mathcal{F}$. Is $\mu_\cdot$ now a homomorphism? We claim so. Proof: Fix positive ϵ, and let V be the 0-neighborhood in $SF(X, \mathcal{M}, M)$ consisting of all $\tilde{\mu}$ with $|\tilde{\mu}(A)| \leq \epsilon$ for every measurable A. Let U be the 0-neighborhood in $\mathcal{F}$ consisting of all f with $\int_X |f| \leq \epsilon$. Then, for any f in U and any measurable A, $\int_A |f| \leq \epsilon$, whence $|\int_A f| \leq \epsilon$. That is, for any f in U and any measurable A, $|\mu_f(A)| \leq \epsilon$. That is, for any f in U, μ_f is in V. That is, $U \subset \mu_\cdot^{-1}[V]$. We conclude that $\mu_\cdot^{-1}[V]$ is a 0-neighborhood in $\mathcal{F}$. Hence, $\mu_\cdot$ is indeed a homomorphism. [To see in more detail why we choose on $\mathcal{F}$ the 0-neighborhoods we did, one

can (easily) verify the following fact: our 0-neighborhoods in $\mathcal{F}$ are precisely those subsets of $\mathcal{F}$ which contain as subsets sets of the form $\mu_\cdot^{-1}[V]$ for V some 0-neighborhood in $SF(X, \mathcal{M}, M)$. In other words, we secretly used the mapping $\mu_\cdot$ to discover our 0-neighborhoods in $\mathcal{F}$.]

To summarize, we so far have homomorphism $\mu_\cdot$ from abelian topological group $\mathcal{F}$ to complete abelian topological group $SF(X, \mathcal{M}, M)$. We claim, finally, that $\mathcal{F}$ is in fact complete. Proof: Let $f_1, f_2, \ldots$ be a Cauchy sequence in $\mathcal{F}$. That is, for every positive ϵ, there is a number N such that $\int_X |f_n - f_{n'}| \leq \epsilon$ whenever $n \geq N$ and $n' \geq N$. Choose $\epsilon = 1/2$ and find N_1; choose $\epsilon = 1/4$ and find $N_2 > N_1$; choose $\epsilon = 1/8$ and find $N_3 > N_2$. Set $f'_1 = f_{N_1}$, $f'_2 = f_{N_2}$, etc. It suffices to show that the subsequence $f'_1, f'_2, \ldots$ of our original sequence converges, in $\mathcal{F}$, to some f. We have by construction: whenever $n \geq N$ and $n' \geq N$, $\int_X |f'_n - f'_{n'}| \leq 1/2^N$. Hence, $\int_X |f'_1 - f'_2| + \int_X |f'_2 - f'_3| + \ldots \leq 1/2 + 1/4 + \ldots = 1$. For each positive number p, denote by A_p the subset of X consisting of all x with $\sum_{n=1}^{\infty} |f'_n(x) - f'_{n+1}(x)| \geq p$. Then $A_1 \supset A_2 \supset \ldots$, while $\cap A_p$ has measure zero (for where $\mu(\cap A_p)$ nonzero, it would contradict the equation just obtained). We conclude that, for almost all x, $\sum_{1}^{\infty} |f'_n(x) - f'_{n+1}(x)|$ is finite. Hence, for almost all x, $\sum_{1}^{\infty} (f'_n(x) - f'_{n+1}(x))$ is finite. But $\sum_{1}^{N} (f'_n(x) - f'_{n+1}(x)) = f'_1(x) - f'_N(x)$. We conclude that, for almost all x, $\lim_{n\to\infty} f_n(x)$ exists. By modifying each f_n on this set of measure zero (a modification which makes no difference in $\mathcal{F}$), we have function f such that $f(x) = \lim f_n(x)$ for all x. Thus, $\lim |f(x) - f_n(x)| = 0$ for all x, whence, by Theorem 16, $\lim \int_X |f - f_n| = 0$. Thus, our sequence in $\mathcal{F}$ converges to f.

We can summarize all this as follows:

Theorem 17. The mapping $\mu_\cdot$ is a one-to-one homomorphism from the complete abelian topological group $\mathcal{F}$ to the complete abelian topological group $SF(X, \mathcal{M}, M)$.

In a certain sense, Theorem 17 summarizes the basic facts about the Lebesque integral. That $\mathcal{F}$ is a complete abelian topological group describes the properties of measurable functions. That the mapping $\mu_\cdot$ exists describes the existence of the Lebesque integral. That

the mapping is to $SF(X, \mathcal{M}, M)$ describes the dependence of the integral on the region of integration. That the mapping is a one-to-one homomorphism describes the dependence of the integral on the function integrated. Theorem 17 represents a slightly different viewpoint toward the notion of an integral. Interesteis centered, not so much on the value of the integral of a particular function over a particular region, but rather on the general framework in which the notion "integral" lies. For example, one could think of the homomorphism $\mu_\bullet$ as "representing" Lebesque measure in a certain sense. Other measures on $X, \mathcal{M}$ could then be described in terms of other homomorphisms from $\mathcal{F}$ to $SF(X, \mathcal{M}, M)$. Properties of these measures could then be described as properties of the corresponding homomorphisms. Similarly, we need not think of a function as just a rule which assigns, to points of X, real numbers. We could also think of a function as being "represented" by its measure, μ_f. From this point of view, the actual numerical value of f at a single point x is irrelevant (for a set with one point has measure zero): it is the "global behavior" of f that counts.

We give one example of this viewpoint. Is the mapping $\mu_\bullet$ onto, i.e., is every element $\tilde{\mu}$ of $SF(X, \mathcal{M}, M)$ some μ_f for f in $\mathcal{F}$?

Example. Fix a point x of X. Let $\tilde{\mu}$ be the element of $SF(X, \mathcal{M}, M)$ with the following action: $\tilde{\mu}(A) = 1$ if x is in A, and zero otherwise. There can be no f with $\tilde{\mu} = \mu_f$. Indeed, let A be the set with single element x. Then $\tilde{\mu}(A) = 1$, while, for any f, $\int_A f = 0$ (since A has measure zero), whence $\mu_f(A) = 0$.

Thus, $\mu_\bullet$ is not onto. [In fact, although we shall not prove it, the following is true: element $\tilde{\mu}$ of $SF(X, \mathcal{M}, M)$ is some μ_f if and only if $\tilde{\mu}(A) = 0$ for every A with zero Lebesque measure.] The example above suggests that elements $\tilde{\mu}$ of $SF(X, \mathcal{M}, M)$ which are not μ_f's represent "(nonexistant) functions which get infinite on a set of measure zero, in such a way that their integrals over this set is nonzero". Although such functions do not exist as functions, they do "exist" as elements of $SF(X, \mathcal{M}, M)$. Elements of $SF(X, \mathcal{M}, M)$ are sometimes called distributions. By Theorem 17, then, every measurable function gives rise to a distribution. One suspects, correctly, that distributions should be objects of interest.

Finally, we describe a sense in which Theorem 17 makes the Lebesque integral seem to have a more fundamental character than, say, the Riemann integral. Denote by $\mathcal{F}'$ the subset of $\mathcal{F}$ consisting of f with $\int_X |f|$ finite: for simplicity, we restrict consideration to this $\mathcal{F}'$. Next, denote by $\mathcal{S}$ the subset of $\mathcal{F}'$ consisting of the step functions. Now, $\mathcal{F}'$ is a complete abelian topological group, and $\mathcal{S}$ is some subset of $\mathcal{F}'$. It is "obvious" what the integral of a step function should be, and hence we know "a priori" what $\mu_*(\mathfrak{s})$ should be for $\mathfrak{s}$ in $\mathcal{S}$. In what sense do we thereby "know" what $\mu_*(f)$ should be for other f in $\mathcal{F}'$? We claim: given any f in $\mathcal{F}'$, and any f-neighborhood U in $\mathcal{F}'$, there is an s in $\mathcal{S}$ with s in U. Proof: It suffices to find, for any positive ϵ, step function s with $\int_X |f - s| \leq \epsilon$. Find measurable sets $A_1 \subset A_2 \subset \ldots$ with each $\mu(A_i)$ finite, with f bounded on each (A_i), and with $\cup A_i = X$. Then, since $\lim \int_{A_i} |f| = \int_X |f|$ is finite, there is some n with $\int_{X \cap A_n^c} |f| \leq \epsilon$. Approximate f on A_n by step function s with $|f(x) - s(x)| \leq \epsilon$ for every x in A_n, and let $s(x) = 0$ for x not in A_n. Then $\int_X |f - s| \leq \epsilon + \epsilon \mu(A_n)$. Since ϵ is arbitrary, we have proven our claim.

Thus, even in $\mathcal{F}'$, functions can be approximated, in the sense of the 0-neighborhoods of $\mathcal{F}'$, by step functions. Hence, the completion of the abelian topological group $\mathcal{S}$ is precisely the complete abelian topological group $\mathcal{F}'$. Since we know (without any fancy definitions) what μ_* is on $\mathcal{S}$, and since $\mathcal{F}'$ is obtained from $\mathcal{S}$ by completion, we "know already" what μ_* must be on $\mathcal{F}'$. [More precisely, any element f of $\mathcal{F}'$ is a limit of a Cauchy sequence of step functions, $s_1, s_2, \ldots$ Then $\mu_*(s_1), \mu_*(s_2), \ldots$ is a Cauchy sequence in $SF(X, \mathcal{M}, M)$, whence it converges to some element of $SF(X, \mathcal{M}, M)$. This element of $SF(X, \mathcal{M}, M)$ is our choice for $\mu_*(f)$.] Since $\mathcal{F}'$ is already complete, we cannot "learn how to integrate any more functions (i.e., more than are already in $\mathcal{F}'$) in this way". That is to say, Lebesque integration in a sense exhausts what one can do starting from the step functions.

The problems below are a mixed bag: they vary from the concrete to the rather abstract; from easy to very difficult. This selection has been made with the intention of providing people with different inclinations and backgrounds with at least a few interesting questions each. In general, I have tried to make the problems as open-ended as possible, because problems in real life are open-ended.

I would like those registered for the course (and, of course, any others who wish to) to hand in to me a few pages of work on some number (say, three or four) of these problems. I would urge you to select problems according to your background and interests. If, for example, you have had rather little exposure to mathematics, then you should try the easier problems first. In any case, the best problems are those in which some effort must be exerted, and in which some progress is made. Do not feel that it is necessary (or always even desirable) to solve any particular problem. I would prefer that you simply report on what progress has been made: constructions or ideas which have been tried, how they have worked out, what seems to be helpful and what not, interesting counterexamples, etc. In particular, work on a problem may lead off in some other direction: if so, by all means follow it out, and write down where it has led. In particular, if you have any open questions of your own in this general area that you would like to look at instead of these problems, feel free to do so.

I would like to have the work turned in to me, if possible, by Tuesday, November 26. I will go over what you have written and make comments and suggestions.

Problems

1. Let M and M' be complete abelian topological groups, G their direct sum, and φ and φ' the corresponding homomorphisms from M and M', respectively, to G. Prove that, given any complete abelian topological group $\tilde{G}$, and homomorphisms $\tilde{\varphi}$ and $\tilde{\varphi}'$ from M and M', respectively, to $\tilde{G}$, there is a unique homomorphism γ from G to $\tilde{G}$ with $\gamma \circ \varphi = \tilde{\varphi}$ and $\gamma \circ \varphi' = \tilde{\varphi}'$.

2. Define a discrete measure ("all the measure is concentrated at single points"), and an indiscrete one. Prove that every measure can be written in one and only one way as the sum of a discrete measure and an indiscrete one.

3. State and prove: for M, M', and M'' complete abelian topological groups, $M \otimes (M' \oplus M'') = (M \otimes M') \oplus (M \otimes M'')$.

4. Does there, or does there not, exist a measure space $X, \mathcal{M}, M, \mu$ with X and M the reals, with every Lebesque measurable set in $\mathcal{M}$ together with other sets, and with $\mu(A)$ the Lebesque measure of A whenever A is Lebesque measurable (i.e., can one "extend" Lebesque measure)?

5. Let K be any subset of [0,1]. Prove that there exists Lebesque measurable A with $K \subset A$ such that, for every Lebesque measurable B with $K \subset B \subset A$, $\mu(B) = \mu(A)$. Is something like this true for measure spaces in general?

6. Prove that every complex measure can be written in one and only one way as a real measure plus i times another real measure.

7. Define the footprint of measurable A to be the set of all $\mu(B)$ for B a measurable subset of A. Prove that, in Lebesque measure, the footprint of a set of finite measure is closed and bounded. Are there generalizations to a general measure space?

8. Carry out the details of step 5 on page 62.

9. What can be done about that awkward business at the end of the construction of products of measure spaces? For example, are either of the "essential steps" true?

10. Find an example of a measure space which is not sigma-finite, but which is such that X is the union of sigma-finite subsets.

11. Carry out the disjoint union construction for an arbitrary (i.e., not necessarily countable) collection of measure spaces.

Robert Geroch

Topology

1978 Lecture Notes

With 112 Figures

MINKOWSKI
Institute Press

Robert Geroch
Enrico Fermi Institute
University of Chicago

© Robert Geroch 2013
All rights reserved. Published 2013

ISBN: 978-1-927763-17-9 (softcover)
ISBN: 978-1-927763-18-6 (ebook)

Minkowski Institute Press
Montreal, Quebec, Canada
http://minkowskiinstitute.org/mip/

For information on all Minkowski Institute Press publications
visit our website at http://minkowskiinstitute.org/mip/books/

Contents

1 Introduction 1

2 Sets in the Plane 2

3 Distance in the Plane 6

4 Properties of Boundary, Interior, Bounded, and Connected 18

5 Continuous Curves 25

6 Metric Spaces 31

7 Boundary, Interior, Bounded, and Connected in Metric Spaces 38

8 Theorems on Sets in Metric Spaces 48

9 Topological Spaces 51

10 Interior, etc. in Topological Spaces 58

11 An Example: Connected Sets with Boundary Attached 67

12 Compactness 69

13 Continuous Mappings of Topological Spaces 74

14 Applications of Continuous Mappings 82

15 Conclusion 87

Appendix 91

1. Introduction

A key activity in mathematics is the attempt to capture certain intuitive ideas. By "capture", we mean to express these ideas in a form which is precise and clear, i.e., in a form which can be communicated without requiring that one's listener be instilled with all the nuances of one's own intuition. Two closely related activities are those of isolating and generalizing. One wishes to know exactly which features of some collection of mathematical objects are relevant to which conclusions, and how these various features are related in more general contexts.

One example of these activities involves the system of real numbers. One knows that one can, among other things, add numbers, and that this operation satisfies certain properties. [For example, $3+19 = 19+3; (3+19)+31 = 3+(19+31)$.] These properties are, of course, intuitively clear, and indeed represent part of our mental picture of what "addition" means. The response of the mathematician to this state of affairs would be the following. What is really relevant here? The key elements are apparently a set (in this example, the set of real numbers), together with a specified operation on this set (in the example, addition of numbers). The "addition-ness" of this operation would be expressed by demanding certain properties (such as those above) of this operation. One thus "captures" addition of real numbers by introducing the notion of a set together with an operation, subject to certain properties. This "distilled essence of addition" is called an abelian group. One example of an abelian group is of course the numbers under addition. But it turns out that there are many others. For example, the non-zero numbers under multiplication (which operation, of course, has all the same properties as addition, e.g., $2 \times 3 = 3 \times 2; (2 \times 3) \times 7 = 2 \times (3 \times 7)$) also form an abelian group. There are also examples in geometry, for instance, the set of rotation about the origin in the plane, or the set of translations in Euclidean three-dimensional space. An abelian group, then, is just "the numbers under addition", but shorn of all extraneous elements (such as the issue of which numbers are larger than which others; which are written with straight and which with curved lines), and generalized to allow other examples. at least two benefits flow from, in this example, formulating the notion of an abelian group. First, one acquire a sense of a deeper understanding of what "addition" is all about.

The second is more practical. It turns out that many facts about addition of numbers are actually true more generally for any abelian group. Thus, one has only to establish such facts once – for the general abelian group – whence they will automatically be available within each particular example.

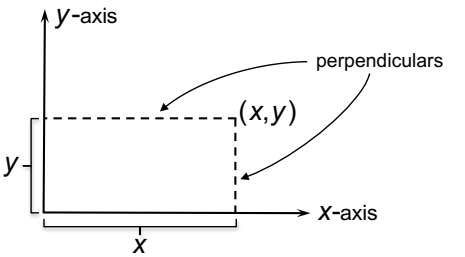
Mathematics proceeds on two levels: the intuitive, in which one forms a mental picture of the objects with which one deals and their relationships with each other, and the more formal, which consists of the definitions, theorems, and proofs. It is the interplay between these two levels – and in particular the translation from the first to the second – which is one of the essential and exciting features of mathematics.

It is our purpose here to see mathematics in action: how this interplay works, how mathematics operates and builds, how mathematicians think. This can apparently only be done by means of an example, and our example is to be the branch of mathematics called topology, Why topology? It has several attractive futures. First, this is an area in which all of us have already a rich and strong intuition, on which we shall be able to draw. Second, topology has the feature that interesting theorems are very near the surface. That is to say, topology has a particularly high ratio: (what comes out as theorems) / (what has to be put in as definitions, etc.). A third attractive feature is that topology is very different from the mathematics with which one may be familiar, e.g., algebra, and so one is provided the opportunity to see the breadth that is mathematics. Finally, topology is attractive because it is an active field of current research in mathematics.

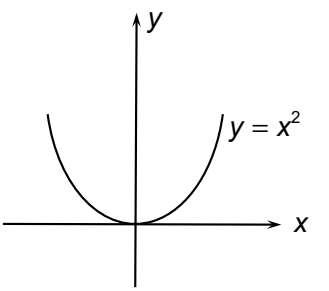
2. Sets in the Plane

In this section we do two things. First, we shall try to indicate, by means of some examples, the kind of structure one is (and the kind one is not) trying to capture in topology. That is, we indicate what topology is all about. Second, we shall discuss the intuitive meaning of certain notions which are, ultimately, to be defined, i.e., to be made precise in topology. Both of these goals are to be accomplished by a discussion and description of certain sets in the plane.

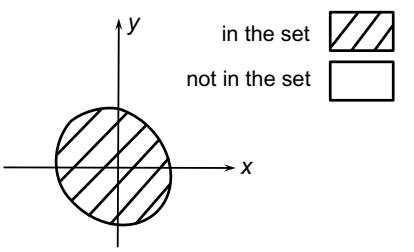
Recall the Euclidean plane, i.e., the space of plane geometry. We may represent points in the plane by the usual Euclidean coordinates, x and y. That is, we introduce orthogonal $x-$ and $y-$axis, as shown on the right. Than any point in the plane is represented, by two numbers, its $x-$ and $y-$coordinates as follows. First, drop perpendicular from the point to the $x-$ and $y-$axes, as shown. Then the $x-$coordinate of the point is the distance from the origin to the foot of the perpendicular along the $x-$axis, and similarly for the $y-$coordinate. The point in the figure, then, would be represented (x, y). Any point in the plane is represented in this way by two numbers, and, conversely, any two numbers represent a unique point of the plane.



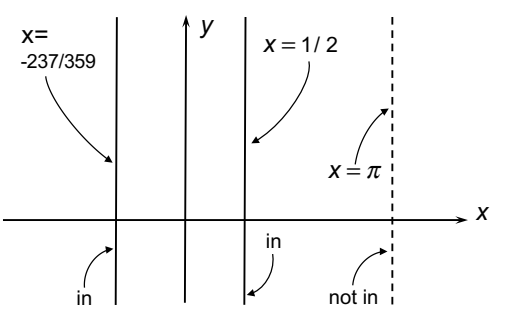
We shall be interested, for the moment, in various sets in the plane. To characterize a set, one must give some clear, unambiguous prescription for deciding which points of the plane are in the set and which are not in the set. This can be done in a number of different ways. For example, one could give an equation in the coordinates (x, y), e.g., "the set of all points (x, y) in the plane for which $y = x^2$". This is, of course, such a prescription: To decide whether or not a given point, represented (x, y), is in this set, one computes x^2, and checks to see whether or not it is equal to y. The resulting set is of course that represented by the parabola in the figure.



One could also specify a set by giving an inequality in the coordinates, e.g., "the set of all points (x, y) for which $x^2 + y^2 < 1$". This set is of course the disk shown. More generally, any combination of equations and inequalities on (x, y) will provide a clear, unambiguous prescription, and thus will define a set in the plane.



For example, "the set of all points (x, y) for which $x^{11} - 13/6\, x^3\, y \leq .16\, y^5$ or $y^3 - xy \neq 0$, except that all points with $y = 3$ are excluded" describes a set in the plane, for again, given a candidate, (x, y), one could decide whether or not it is in the set. Sets can also be specified by other than algebraic equations. For instance, "the set of all points



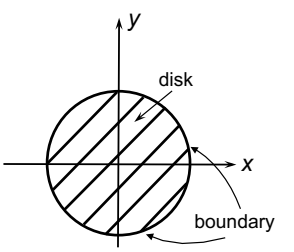
(x, y) for which x is a rational number"
describes a set. This set, for example, consists of a collection of vertical lines in the plane, one for each rational $x-$value. Thus, there are vertical lines at $x = 1/2$ and at $x = -237/359$, but none at $x = \pi$, for π is of course not rational. As other examples, "the set of all points (x, y) for which the decimal expansion of y contains no digit 7" or "the set of all points (x, y) for which the area of that part of the circle of radius x^2 and center (x, y) which lies to the left of the $y-$axis is less than 3" specify sets in the plane.

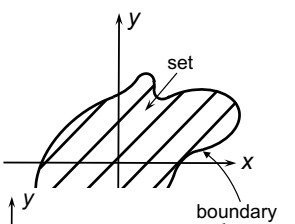
The following are examples of things which would not be regarded as specifying sets in the plane: "If you want to know whether (x, y) is in the set, flip a coin: heads, in; tails, out." "Point (x, y) is in the set if y is a positive integer, and the Bears win Super Bowl y." These would be regarded as unclear and/or ambiguous.

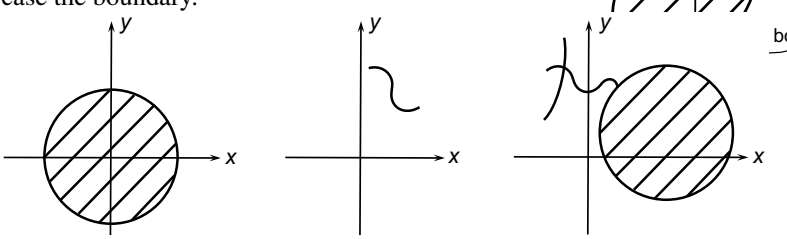
I hope these few examples make the distinction clear. In fact, there is an active branch of mathematics which deals in part with the issue of what, precisely, "specify a set" should mean.

We next wish to discuss certain intuitive features of various sets in the plane. We emphasize that the purpose of this discussion is merely to develop some common ground for later use. What follows is *not* mathematics.

Given a set in the plane, we wish to mean by its boundary the set of all points "on the edge of the given set". Consider, for example, the disk, the set of points (x, y) with $x^2 + y^2 < 1$. Its boundary should be circle shown, i.e., the set of points (x, y) with $x^2 + y^2 = 1$. This circle is also the boundary of the "disk with its edge included", i.e., of the set of points (x, y) with $(x^2 + y^2 \leq 1$. For the more complicated set illustrated in the second figure, its boundary would again be its edge, as indicated. Given a set in the plane, we wish to mean by its interior the set of all points which are "actually inside the set, i.e., neither external to the set nor on its boundary". Thus, for the two examples illustrated above, the interiors are to be the cross-hatched regions, excluding in each case the boundary.

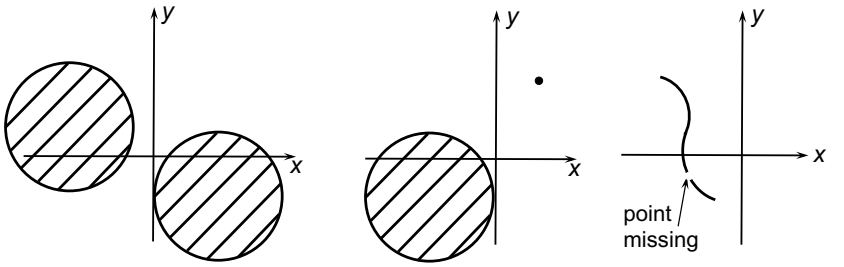


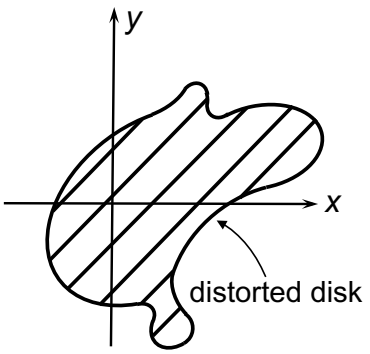




Thus, we have two intuitive features of sets, the boundary and interior. Each is to assign to a given set another set which is its interior. Each is to assign to a given set another set (i.e., to a given set the set which is its boundary), or the set which is its interior). The next two features will have a somewhat different character: They will be properties which a given set is to either have or not have. We wish to describe a set as being connected if it "consists of just one piece, not of two or more". As examples, the sets illustrated above (a disc, a segment of a curve, a disk with a segment of a curve going off of it and another segment of a curve crossing it) are to be regarded as connected. By contrast, the sets below (two disks, a disc and a point, a curve with a point missing) are not to be regarded as connected, for each "consists of two or more pieces".

Finally, we wish to regard a set as bounded (nothing to do with "boundary") if it "remains in a finite region of the plane; does not go off to infinity". As examples, all of the sets illustrated on the previous page were bounded. Some examples of sets which are not to be regarded as bounded are the entire plan ("the set of all points (x, y)"), and an infinitive straight line (for example, "the set of all (x, y) with $x = 1$").





The four features described above, boundary, interior, connected, and bounded, will, as it turns out, all be incorporated into topology. By way of indicating in a general sense what topology is about, we can also give some examples of "features of sets" which will have no place in topology. Such features include "being a disk", "being a straight line", "being more than one foot in size". What is it about these features which will make them "non-topological"? It is a distinction which is difficult to make clearly at this point. All of these features refer to how the points of the plane are arranged to form the plane. The essence of the distinction is the following. Imagine that our sets are drawn, not on the rigid plane we have been imagining, but rather on a rubber sheet. We now allow our rubber sheet to be stretched and pulled (but not folded or torn), and ask which of our features thereby remain the same. For example, "being connected" does remain the same. A disk, for instance, is connected. If we were to draw our disk on a rubber sheet, and then stretch and pull, there might result (when the stretched and pulled sheet is imprinted back on the rigid plane) the set illustrated on the right. But this set (a "distorted disk") is of course also connected.

More generally, one convinces oneself that such a distortion (by imprinting on a rubber sheet, stretching and pulling, and then imprinting back onto a rigid plane) always leaves any connected set connected, and any set not connected not connected. In a similar way, the boundary remains the same under our distortion. What "remains the same" means, in this case, is the following. Fix a set (say, a disk), and determine its boundary (in this example, its circular edge). Now imprint both the set and its boundary (say, with the set in blue and the boundary in orange, to keep them separate) on a rubber sheet. Stretch and pull the sheet, and then imprint both colors back onto the rigid plane. Then, as seems clear, the boundary of the resulting blue set will be just the resulting orange set. Similarly, interior and bounded remain the same under this stretching and pulling. But note, by contrast, that our non-topological properties are altered by this operation. That is to say, a set which is a disk, which is a straight line, or which is less than one foot in size, need not remain so under stretching and pulling of the rubber sheet. We thus have at least a rough criterion for what is likely to be a topological feature of sets in the plane and what is not. Topology is sometimes called "rubber sheet geometry".

Two other possible features of sets which will play no role in topology are "being excessively large" and "being attractive". For these, one does not perhaps have a sufficiently clear intuitive notion – or at least none which is likely to be in reasonable agreement with others – to even get started. It would be very difficult to pin down what it is one is trying to capture.

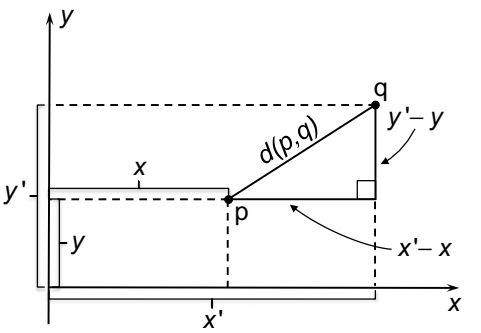
So, we have decided to focus on four features of set in the plane: boundary, interior, connected, and bounded. Why did we call our treatment of these features an "intuitive discussion"? Why did we not instead regard "The boundary of a set is its edge.", "The interior of a set consists of the points inside the set.", "A set is connected if it consists of no more than one piece.", and "A set is bounded if it remains in a finite region of the plane." as the *definitions* of these four terms? The reason is that these sentences are

not sufficiently precise to be regarded as proper definitions in mathematics. A definition in mathematics is supposed to be unambiguous, i.e., such that someone, on reading the definition but without using subtle nuances of what words mean in English, can decide in every case whether or not the conditions of the definition are applicable. But our characterizations of "boundary", "interior", "connected", and "bounded" are ambiguous. Some examples will illustrate this point. Consider the set in the plane consisting of the single point (0, 0), i.e., the point at the origin. What is the boundary of this set? One may have an opinion on this question (presumably, either that the boundary consists of no points, or that the boundary is just the single point at the origin), but I would venture that one is not so confident that others will share that opinion as one was, say, with the disk. What is the interior of this set? Is it this one point or no points. Consider, as a more extreme example, the set illustrated in the second figure on page 4. What is the boundary of this set? Is it all points in the plane? Just those (x, y) with x rational? Those with x irrational? None of the points in the plane? What is the interior of this set? Is this set connected? Clearly, our discussion of the terms "boundary", "interior", "connected", and "bounded" did not provide a clear prescription for deciding what the boundary or interior of a set is, whether a set is connected or bounded. Our discussion, in short, was not a definition in the sense of mathematics.

We see that a definition is a somewhat different thing in mathematics than in everyday English. [It is too bad that a single word is used for both.] For example, my dictionary defines "chair" as "a seat, usually movable, for one person". Clearly, this definition is intended merely to give one a general idea of what sort of objects are conventionally called chairs and what sort are not. It is certainly not intended to permit one to make, unambiguously for every object one sees throughout one's life, the decision as to whether or not that object is a "chair". In particular, "usually ..." could never be inserted in a definition in mathematics. We shall see some examples of mathematical definitions shortly.

3. Distance in the Plane

We now have before us a typical example of the "raw material" of mathematics. We have some intuitive notions (namely, that of the boundary and interior of a set; that of whether a set is connected or bounded), and also a general idea of the limits of our intuition. The task we face, as mathematicians, is to "capture" these notions, i,e,. to formulate some precise definitions, within mathematics, of the words "boundary", "interior", connected", and "bounded". We remark that we are not exactly trying to "discover" the definitions (as though they already exist, somewhere out there, and we are merely trying to bring them to paper) – but rather we must "concoct" them. The only requirements on our concocted definitions are, first, that they be clear and unambiguous, and, second, that they seem to reflect our intuitive ideas of what these words mean. In this section, we make a preliminary attempt at such definitions. This attempt will be refined shortly.



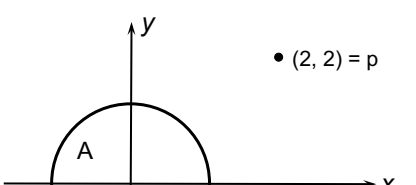
The first decision we face – a critical one – is that of what structure we shall use in the formulation of our definitions. The plane, of course, has a great deal of structure: that of "straightness" of lines, of "orthogonality" of intersecting straight lines, of "circle-ness" of circles, etc. But presumably only a small part of all this structure is relevant to the sorts of things we are after. What we want to do, then, is somehow isolate what it is about the plane which gives rise to our ideas about "boundary", etc. In practice, one would normally try several different structures, selecting them initially based on one's judgment as to what looks promising, and finally based on what choices actually lead to appropriate definitions. [How accurate one's judgment is on such issues is one of the important distinctions between a good and a poor mathematician.] We, however, do not have the time to explore all of these false starts (e.g., to try "straightness", decide that it seems very difficult to formulate a definition for "boundary" in terms of straightness, and then finally realize that straightness was not a very good choice in the first place, since of course stretching and pulling of the rubber sheet certainly destroys straightness). We, therefore, merely jump to that structure which, as it turns out, works. Recall the notion of the geometrical distance (e.g., as measured by a ruler) between two points in the plane. We denote the distance between points p and q by $d\,(p,\,q)$, There is, of course, a simple formula for this distance in terms of the coordinates for the points. Let p have coordinates $(x,\,y)$, and q coordinates $(x',\,y')$. Then, as shown in the figure, one constructs a right triangle, one leg of which has length $(x'-x)$, the other, length $(y'-y)$. But the hypotenuse of this triangle is the line segment joining p and q – a segment whose length we have agreed to denote $d\,(p,\,q)$. So, by the Pythagorean theorem, we have $d\,(p,\,q)^2 = (x'-x)^2 + (y'-y)^2$. Taking the square root of both sides, we obtain our formula for $d\,(p,\,q)$:

$$d\,(p,q) = \sqrt{(x'-x)^2 + (y'-y)^2}$$

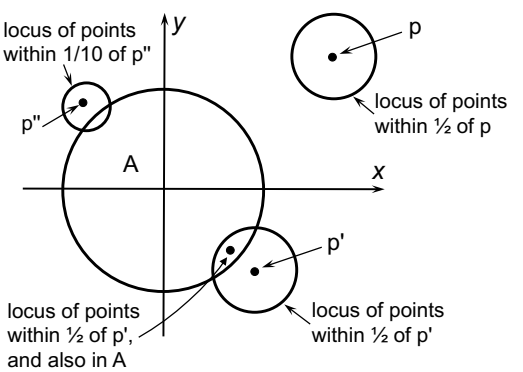
Thus, for example, the distance between the point labeled (1, 2) and the point labeled (4, 6) is $\sqrt{(4-1)^2+(6-2)^2} = \sqrt{3^2+4^2} = \sqrt{9+16} = \sqrt{25} = 5$.

Our program. then, is to give expression to our ideas about "boundary", etc., using this distance.

We begin with the notion "boundary". Let us, to fix ideas, focus on one particular set A, say the unit disk centered at the origin, i.e., the set of



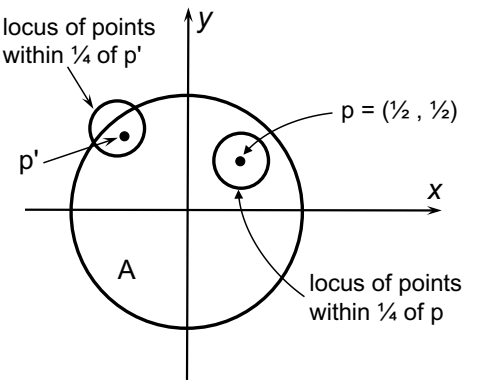
all points (x, y) with $x^2 + y^2 < 1$. Consider some point p, such as the point $(2, 2)$, as shown in the figure. Now we clearly do not want this point p to be in the boundary of our set A. The question is: What is the basis for this attitude toward the point p? One might answer by saying that the distance of the point p from the set A is greater than zero. But this, unfortunately, is not a very useful expression of our attitude, for we do not know what "distance of a point from a set" means (only "distance of a point from a point"). Alternatively, one might say that the distance of the point p from the point at the origin (a distance, in this example, of about 2.8) is greater than 1, the radius of the disk A. Although this statement certainly does say what we want to say for this particular example, it is, unfortunately, rather strongly tied to the example itself (i.e., to the fact that A is a disk). If, for example, A were some more complicated set, then we would not be able to say"... the radius of the disk A. What we want to say, of course, is that our point p should not be on the boundary of the disk because it is "well away" from the disk. After a few more similar tries, one might hit upon the following statement: There is no point within distance $1/2$ of p which is also in A. This statement is, of course true. Indeed, the locus of points within distance $1/2$ of p is just the disk of radius $1/2$ centered at p, and this disk does not intersect the set A, i.e., it has no points in common with A. So, one might be tempted to regard a point p as "well outside" of set A (so, not on the boundary of A) if there is no point within distance $1/2$ of p which is also in A. But this, unfortunately, is still not quite right.



Consider, for example, the point p' of the figure. Now this p' is certainly "well outside" the disk (perhaps not as "well outside" as was p, but at least "well enough outside" so that we do not want p' to be on the boundary of A). However, there *is* a point within distance $1/2$ of p' which is also in A, namely the point indicated in the figure. So, one might think of changing our statement to: Point p is well outside of set A if there is no point within distance $1/10$ of p which is also in A. This statement of course holds for both our point p and our point p'. But this will not work either. Consider the point p'' in the figure. It is still to be "well outside" of A, but now there *is* a point within distance $1/10$ of p'' and also in A. Our mistake is that we are committing ourselves to "how well outside" (i.e., to the distance, $1/2$ or $1/10$) *before* we know which point (p or p' or p'') we are to consider. All we have to do is reverse the order of commitment (i.e., allow ourselves to choose the distance *after* we know the point to be considered). This we do by the following statement: We regard a point p as "well outside" of A if, for some positive number ϵ, no point is within ϵ of p and also in A. [Note the word "some": We get to choose ϵ, the distance, after we know what p is.] This statement works. Consider, for example, the figure above. The point p, according to this statement, is well outside of A, for we may choose $\epsilon = 1/2$: There is no point within $1/2$ of p which is also in A. That is, there is *some* positive ϵ (namely, $1/2$) such that no point within ϵ of p is also in A. The point p' is also "well outside" of A, for *some* positive ϵ (namely, $1/10$), there is no point within ϵ of p' which is also in A. Finally, the point p'' is also "well outside" of A, e.g., by choosing $\epsilon = 1/100$.

All we have done so far is identify certain points which we do not want to be in the boundary of our set, namely those which are "well away" from the set, in the sense above. But we also want to exclude from the boundary those which are "well inside" the set. Consider, for example, the point p in the figure, given, say by $(1/2, 1/2)$. This

p, of course, is not to be on our boundary. How do we express the fact that it is "well inside" our set A? Again, one would try various criteria, most of which would not look too promising. [E.g., the distance of p from the origin (about .7 in this example) is less than the radius of the disk. But this criterion depends on our particular set A.]



Finally, one would notice that the "well inside-ness" of p can be expressed thus: Every point within distance $1/4$ of p is also in A. Indeed, the locus of points within distance $1/4$ of p is the small disk shown, and all such points are in our set A. Again, however, the criterion "all points within distance $1/4$ of p be also in A" turns out not to be quite right. The point p' of the figure, which we would certainly wish to regard as being "well inside" A, does not have the property that all points within distance $1/4$ of p' are also in A. Again, we must allow ourselves to choose "how well within", i.e., to choose the distance, *after* we know which point we are to consider. Thus, as before, end up with the following: We regard point p as "well inside" set A if, for some positive ϵ, every point within ϵ of p is also in A. With this criterion, then, the point p in the figure above is "well within" A (choosing $\epsilon = 1/4$), as is the point p' (now choosing, say $\epsilon = 1/100$).

We now have all the ingredients to formulate our definition. We want the boundary of A to be what is left over after one excludes all the points which are "well outside" A, in the sense above, and also all points which are "well inside" A, in the sense above. All we have to do is figure out how to say all this in one sentence. One convinces oneself (e.g., by reexamination of the figure on page 7) that point p fails to be "well outside" of A provided that, for every positive ϵ, there is a point within ϵ of p and also in A; and that point p fails to be "well inside" A provided that, for every positive ϵ, there is a point within ϵ of p and also outside of A. But the boundary of A is supposed to be just those points which both fail to be "well outside" of A and fail to be "well inside" A. Thus, we arrive at:

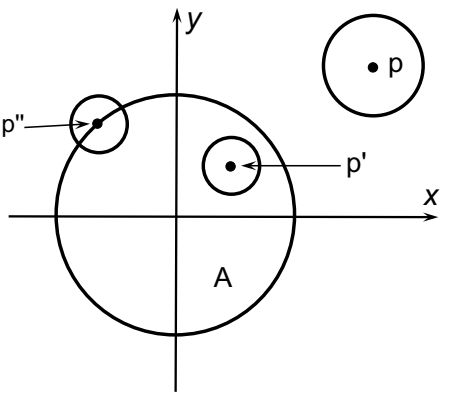
Definition. Let A be a set in the plane. Then the *boundary* of A is the set of all points p such that, for every positive number ϵ, there is a point q with $d(p, q) \leq$ *epsilon* and q in A, and also a point q' with $d(p, q') \leq \epsilon$ and q' outside of A.

A few remarks are in order. First, the words which follow "Definition" above are the first things we have written which may be regarded as "real mathematics". All else – all the discussion, false attempts, examples, and so on – was merely some general remarks to motivate and get one in the mood. What preceded the definition, on the other hand, is a good example of the sort of activity in which mathematicians are engaged. [Only a small fraction of the time of a mathematician is normally spent "calculating", or "working with formulae".] Good judgment is no less important an attribute for a mathematician than for anyone else. The second remark involves again the issue of what a definition is supposed to do in mathematics. The role of this definition is to permit one to decide, unambiguously, which points are to be in the boundary of any given set A and which are not. This is quite a different role from that of the definition of "chair". To decide whether or not an object is a "chair", one may use the dictionary definition as a rough guide, but one also uses one's lifetime of experience with chairs. The dictionary definition is not intended to replace that experience. But not so in mathematics. The following point, which sounds trivial when stated, but which seems often to be overlooked in the heat of the moment, cannot be overemphasized: *To decide what the boundary of a set is, one uses precisely what it says in the definition, nothing more and nothing less*. One does not use one's ideas or thoughts about boundaries; one does not think. Indeed, this is why we had to work so hard to get the definition. This is one feature which sets mathematics apart

from other disciplines.

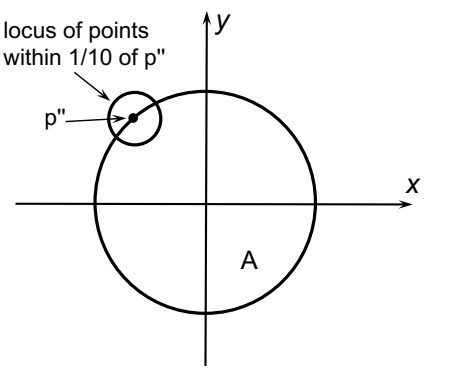
We shall denote the boundary of set A by bnd (A).

We now give some examples of application of the definition to determine the boundaries of various sets. Let us consider first the set we have just used as our example, the disk. Is the point p in bnd (A)? Is it true that, for every positive ϵ, there is a point q within ϵ of p and in A, and also a point q' within ϵ of p and outside of A? The answer is again no. Set $\epsilon = 1/2$. Then there is no point q within distance ϵ of p and in A. Indeed, the locus of points within $1/2$ of p is the disk shown. Is p' in bnd (A)? Is it true that for every positive ϵ, there is a point q within ϵ of p and in A, and also a point q' within ϵ of p and outside of A? The answer is again no. Set $\epsilon = 1/4$. Then there is no point q within distance ϵ of p and outside of A, for the disk shown lies entirely within A. Finally, we consider the point p''. Is p'' in bnd (A)? Is it true that, for every positive ϵ, there is a point q within ϵ of p and in A, and also a point q' within ϵ of p and outside of A?

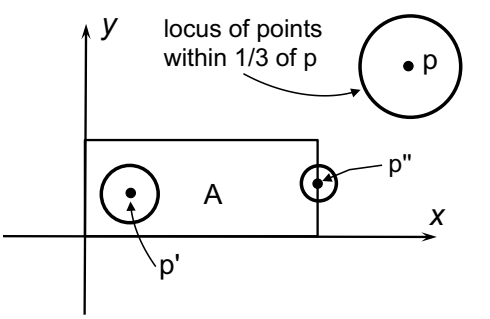


The answer, we claim, is yes. Consider, for example, $\epsilon = 1/2$. The locus of points within distance $1/2$ of p'' is the disk shown. Is there within this disk a point in A, and also in this disk a point outside of A? There certainly is, namely the two points shown.

Consider, then, another value for ϵ, say $\epsilon = 1/10$. Then the locus of points within distance $1/10$ of p'' would be a smaller disk. Is there within this disk a point in A, and also in this disk a point outside of A? Again, there is. Clearly, then, for *every* positive ϵ, there is a point within ϵ of p'' and in A, and also a point within ϵ of p'' and outside of A. That is to say (recalling the definition), the point p'' is on the boundary of A. It should now be clear that bnd (A), the boundary of A, is just the circular "edge" of A, that is, the set of all points (x, y) with $x^2 + y^2 = 1$.

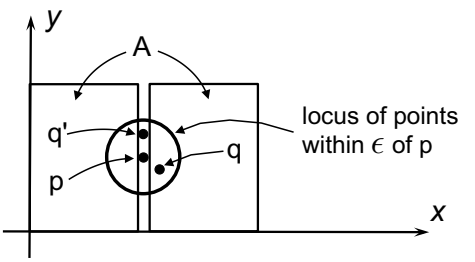


Suppose next that we let A instead be the disk with its boundary, i.e., the set of all points (x, y) with $x^2 + y^2 \leq 1$. Repeating the previous paragraph word for word, we find that bnd (A) is the same as the set above, namely the set of all points (x, y) of the plane with $x^2 + y^2 = 1$.

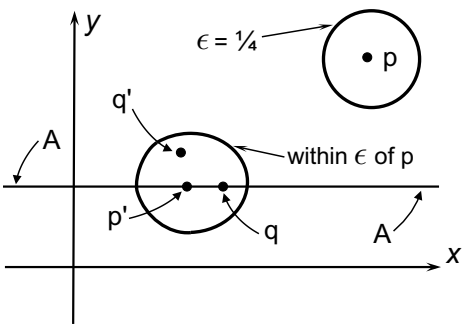


Next, let A be the rectangle in the plane given by all (x, y) with $0 < x < 2$ and $0 < y < 1$. We determine bnd (A). Just as above, the point p is not in bnd (A) (for it is not true that, for every positive ϵ, there is a point within ϵ of p and in A, and also a point within ϵ of p and not in A. Specifically, choose $\epsilon = 1/3$ as shown. Then no point within $1/3$ of p is in A.) Similarly, the point p' is not in bnd (A). The point p'', however, is in bnd (A), for, given any positive ϵ, there is indeed a point within ϵ of p'' and in A, and also a point within ϵ of p'' and outside of A. Thus, the boundary of this set A is again its "edge". More precisely, bnd (A) is the set of all points (x, y) satisfying either i) $x = 0$

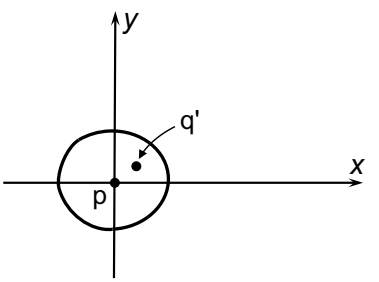
and $0 \le y \le 1$, or ii) $x = 2$ and $0 \le y \le 1$, or iii) $y = 0$ and $0 \le x \le 2$, or iv) $y = 1$ and $0 \le x \le 2$. [These four choices define the four "sides" of the rectangle.]



We now modify this last example slightly. Let A now be "two squares side by side, but without the line between them included", i.e., the set of all (x, y) with $0 < y < 1$ and either $0 < x < 1$ or $1 < x < 2$. Is the point p shown (e.g., $(1, 1/2)$) in bnd (A)? Is it true that, given any positive ϵ, there is a point within ϵ of p and in A, and also a point within ϵ of p and outside of A? The answer is yes. The locus of all points within ϵ of p, for a typical ϵ, is shown. There is indeed a point (namely, q in the figure) within ϵ of p and in A, and also a point (namely, q' in the figure) within ϵ of p and outside of A. Clearly, this holds for any ϵ. So, our point p is in bnd (A).



We consider next a somewhat more exotic example. Let A be the line given by: the set of all (x, y) with $y = 1$. We determine bnd (A). The point p, for example, is not in bnd (A), for it is not true that, for any positive ϵ, there is a point within ϵ of p and in A, and also a point within ϵ of p and outside of A. Choose, say, $\epsilon = 1/4$, as shown. Then *no* point within $1/4$ of p is in A. The point p', by contrast, is in bnd (A). Given any ϵ, choose q (within ϵ of p and in A), and q' (within ϵ of p and outside of A) as shown. We conclude, then, that the boundary of this set A is just the set A itself.



Let A be the single point at the origin, (0, 0). Is this point itself in bnd (A)? It is. Is it true that, given any positive ϵ, there is a point within ϵ of p (the origin) and in A, and also a point within ϵ of p and outside of A? There certainly is: For the point within ϵ of p and in A, choose p itself (clearly within ϵ of p, and certainly in A); for the point within ϵ of p and outside A, choose the point indicated q' in the figure. Thus, for this set A, we have bnd $(A) = A$.

Let A be entire plane, i.e., the set of all points (x, y) in the plane. What is bnd (A)? Let's pick a point p, say the origin, and try it. Is it true that, given any positive ϵ, there is a point within ϵ of p and in A, and also a point within ϵ of p and outside of A? There is not, and in fact there is no point in the plane "outside of A", for now A is the entire plane. So, the origin p is not in bnd (A). Similarly, any other point p we choose will not be in bnd (A). Thus, in this example bnd (A) is the set having no points. [It is customary to express our conclusion in this way, and *not* "There is no bnd (A).] The set having no points, i.e., the set such that for all (x, y) this point is not in the set, is called the *empty set*. [Note that this is an unambiguous statement of what is in and what is out of the set, namely, everything is out.] Our conclusion, then, is that the boundary of the entire plane is the empty set.

Let A be the empty set. What is bnd (A)? Pick a point p, say the origin (0, 0). Is it true that, given any positive ϵ, there is a point within ϵ of p and in A, and also a point within ϵ of p and outside of A? It is not true, and in fact there is no point whatever in A", since now A is the empty set. So, this p is not in bnd (A) – and clearly no other point will be in bnd (A) either. We conclude, then, that the boundary of the empty set is the empty

set.

These examples are intended, not only to illustrate what our definition of the boundary gives under various circumstances, but also to make the following point: When confronted with a definition in mathematics, one should, in the first instance, apply it mechanically to various examples, and see what it gives. Eventually, after one has "experimented" with enough examples, one acquires a feeling for how the definition works, and in particular one becomes able to apply the definition without going through all the words every time. If you do not feel that at this point you can do this for "boundary", I would suggest that you make up some sets of your own, and apply the definition to determine their boundaries.

We next turn to the definition of the second of our notions, that of the interior. Recall that we would like the interior of a set to consist of the points "well inside" that. But we have already obtained, on page 7, a precise formulation of this intuitive notion. We are thus led immediately to:

Definition. Let A be a set in the plane. Then the *interior* of A is the set of all points p such that, for some positive number ϵ, every point q with $d(p, q) \leq \epsilon$ is in A.

Again, we apply the definition to various examples. We shall denote the interior of set A by int (A).

Let A be the disk, as shown in the figure on page 9. Is the point p in int (A)? Is it true that, for some positive ϵ, every point within ϵ of p is in A? The answer is no: Indeed, for any positive ϵ, the point p itself is certainly within ϵ of p, but is not in the set A. Is p' in int (A)? Is it true that, for some positive ϵ, every point within ϵ of p' is in A? Now, the answer is yes. Indeed, choose $\epsilon = 1/4$ (certainly *some* positive ϵ). Then the locus of all points within $1/4$ of p' is the disk shown in the figure, and clearly every point within this disk (i.e., every point within ϵ of p') is in A. So, this p' is in int (A). Is p'' in int (A)? Is it true that, for some positive ϵ, every point within ϵ of p'' is in A? The answer is now no. The locus of points within ϵ of p'', for a typical ϵ, is shown in the figure. But there is, within this small disk about p'', a point not in A, namely the point indicated in the figure. Clearly, this will hold for every ϵ. So, it is not true that, for some ϵ, every point within ϵ of p'' is in A. So, p'' is not in int (A). We conclude, then, that int (A) is just the disk itself, i.e., the set of all points (x, y) within $x^2 + y^2 < 1$. Note that this is of course the answer we wanted, i.e., that our definition seems to do what it is supposed to.

The interior of the rectangle A shown in the second figure on page 9 is just A itself.

Consider next the "two squares" shown in the last figure on page 9. Is p in int (A)? Is it true that, for some positive ϵ, all points within ϵ of p are in A? The locus of points within ϵ of p, for a typical ϵ, is shown in the figure. But it is not true that every point in this disk is in A, for example, the point q' shown in the figure. So, p is not in int (A).

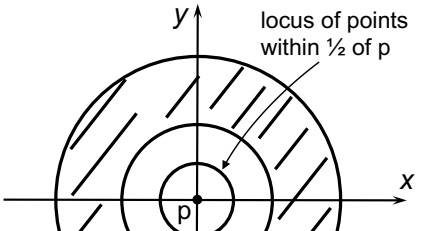
Next, let A be the line as in the first figure on page 10. Clearly, p is not in int (A). What about p'? It is not either. Indeed, for any ϵ (a typical one shown in the figure), there is a point, such as the q' shown, which is written ϵ of p but which is not in A. So, p' is not in int (A). Clearly, then, there are no points in int (A). That is to say, int (A) is the empty set.

For A the single point, as in the second figure on page 10, int (A) is also the empty set.

What is the interior of the entire plane? Consider the point p at the origin. Is it true that, for some positive ϵ, every point within ϵ of p is in A? It sure is: Choose any ϵ you want (say, $\epsilon = 187$), because *every* point in the entire plane is in A. Clearly, the interior of the entire plane is the entire plane.

The interior of the empty set is the empty set.

Finally, let us consider the "disk with a hole in it", i.e., the set A given by all points (x, y) with $x^2 + y^2 < 4$ and $x^2 + y^2 > 1$. [The second condition "removes the hole".] Is the point p at the origin in int (A)? It is not, for it is



not the case that, for every positive ϵ, all points within ϵ of p are in A. Indeed, for $\epsilon = 1/2$, there is not even *any* point within ϵ of p and in A. This example suggests, then, that one should think of the interior as consisting, not so much of the points which are "surrounded" by A, but rather of the points which are "overrun" by A..

We turn next to the intuitive notion "bounded". Recall that bounded is supposed to mean "does not go off to infinity; remains in a finite region". How are we to express this precisely? What we want to say, roughly speaking, is that "the points of A do not keep getting farther and farther away from the 'bulk' of A". After a few false starts, one realizes that this is essentially the same as demanding that "the points of A do not keep getting farther and farther away from some point p of the plane."

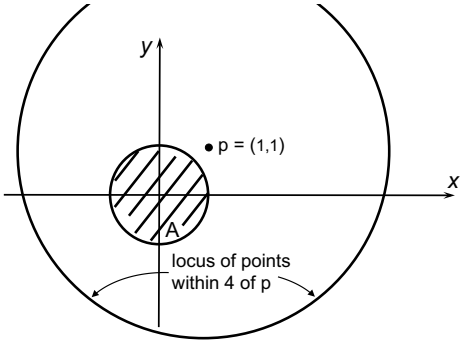
But this idea, finally, can be expressed as follows.

Definition. A set A in the plane is said to be *bounded* if, for some point p of the plane and some positive number c, all points q of A satisfy $d(p, q) \leq c$.

What the definition requires, in other words, is that A lie entirely within some disk (with center p and radius c).

We give some examples.

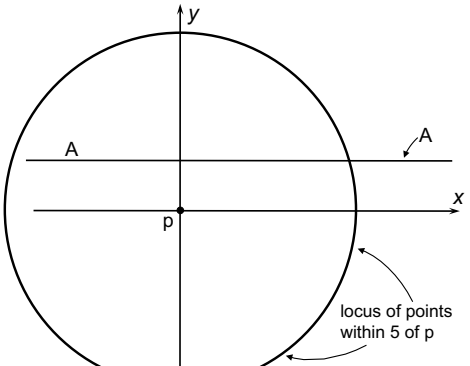
Let A be the usual disk, i.e., the set of all points (x, y) with $x^2 + y^2 < 1$. This set A is bounded. Indeed, choose, say, p the point $(1, 1)$, as shown, and let $c = 4$. The locus of points within distance 4 of p is indicated in the figure. Clearly, all points of A are in this region, i.e., all points of A are within distance ϵ of p. So, A is bounded. [Clearly, many choices of p and c are possible to show that A is bounded. We could, for example, have chosen $p = (0, 0)$, and $c = 1$, or $p = (20, 20)$, and $c = 50$.]



Similarly, the rectangle (the set represented by the second figure on page 9), and the "two squares" (the set represented by the last figure on page 9) are bounded. The set consisting of a single point (say, the origin) is bounded. Choose, say, $p = (0, 0)$, and $c = 738$. Then clearly all points of A (namely the one point which is in A, the origin) is within 738 of p.

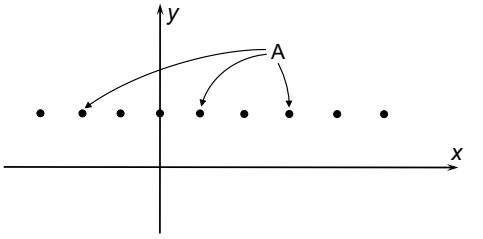
The empty set is bounded. Choose anything you want for p and c (say, p the point $(-13, 4)$ and $c = 2/3$). Then every point in A (the empty set) is within $2/3$ of p, because there are no points in A. The entire plane is not bounded. Is it true that, for some point p and some positive c, all points of A (the entire plane) are within c of p? Hardly, for A is the entire plane, so if all points of A are to be within c of p, then all points of the plane would have to be within c of p. But for no choice of p and c are all points of the plane within c of p. That is, one cannot find p and c such that all points of A are within c of p. That is, our set A, the entire plane, is not bounded.

Let A be the line given by the set of all (x, y) with $y = 1$. Is A bounded? Is it true that, for some point p and positive number c, all points of A are within c of p? Let's try a choice, say p the origin, and $c = 5$. But this doesn't work, for it isn't true that all points of A are within 5 of the origin. Clearly, no p and c will work. We conclude, then that this set A is

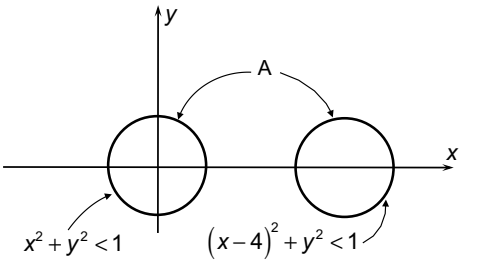


not bounded. [This little discussion isn't quite a full proof of our conclusion. It would not be difficult to give a full proof at this point, but we shall not do so because the result isn't worth the effort. We shall discuss proofs in a moment.]

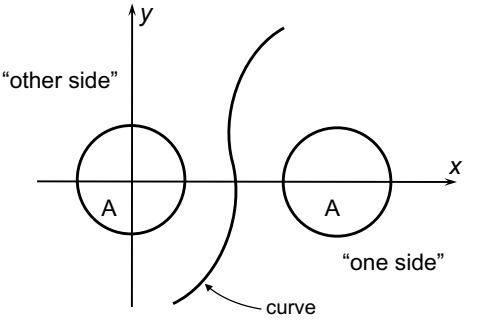
Finally, let A be the set of all (x, y) with $y = 1$ and x an integer. This set is a "bunch of equally spaced points", as shown. This set, clearly, is not bounded.



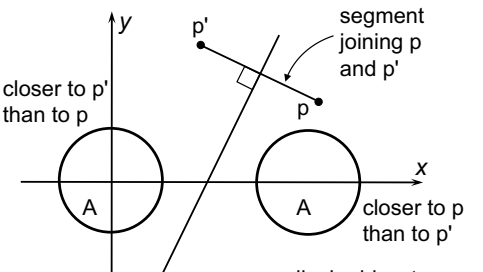
We turn, finally, to our last intuitive notion, that of "connected". [In order to avoid having sentences with many "nots" in them, let us deal with its opposite, "disconnected".] Recall that we wish to regard a set in the plane as disconnected if it "consists of two or more pieces, not joined together". Again, we wish to make this idea precise.



Consider, as an example, the set A consisting of two disks in the plane, as shown. What is it about this set which suggests to us that it consists of "two pieces"? One might say that the essential feature is how A is expressed, in this case, "A is the set of all (x, y) with either $x^2+y^2 < 1$ or $(x-4)^2+y^2<1$". Note the use of "either ... or ...". Is this the tipoff that the set is disconnected? It is a good clue, but this would never serve as a mathematical definition of "disconnected". The problem is that it refers to *how the set is expressed* (in English), and not to *what the set is* (i.e., which points are in and which out).



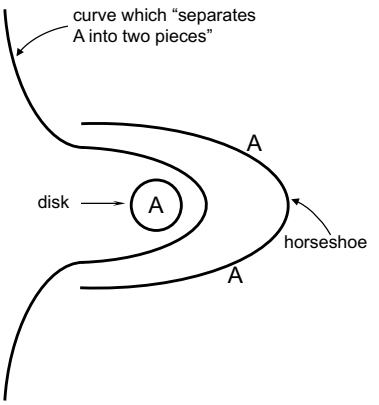
A more promising possibility might be based on the following observation. In the case of our two disks, one can find a curve (such as that shown) such that part of the set is on one side of the curve, part is on the other side, and none of the set meets the curve. This curve, then, makes explicit the feature that the set "consists' of two or more pieces". A definition along these lines, however, involves various other complications. What are we to mean by a "curve" in the plane? What is "one side" of the curve, and the "other side" to mean? That one has a good intuitive idea of what these words mean is not enough, for after all, one also has a good intuitive idea of what "disconnected" means. Perhaps one could resolve these difficulties as follows. Recall, from plane geometry, that, given two points in the plane, the perpendicular bisector of the segment joining those points, the line indicated in the figure, is just the locus of points equidistant from the two points. The points on one side of that perpendicular bisector are nearer to p than to p'; those on the other side nearer to p' than to p.



In this way, one might formulate the following: A set A is disconnected if there exit points p and p' such that some point of A is closer (using our distance) to p than to p', some point of A is closer to p' than to p, and no point of A is equidis-

tant from p and p'. This works, for example, for our set A consisting of two disks. Unfortunately, this is also not a very good definition. The problem now is that it doesn't give the "intuitive answer" for certain examples.

Consider, for instance, the set A consisting of "horseshoe and a disk" as shown. [We do not bother with writing the equations.] We would presumably wish to regard this set as disconnected. But it would not be recognized as disconnected according to the trial definition above. The problem here is that, although this set clearly seems to "consist of two pieces", and although one can, intuitively "separate these pieces by means of a curve", one cannot "separate those pieces by means of a straight line, i.e., by means of a perpendicular bisector". [No straight line can "wander inside the horseshoe to have the disk on the other side of the line from the horseshoe".] The problem with this definition, then, is that it doesn't properly reflect our intuition.

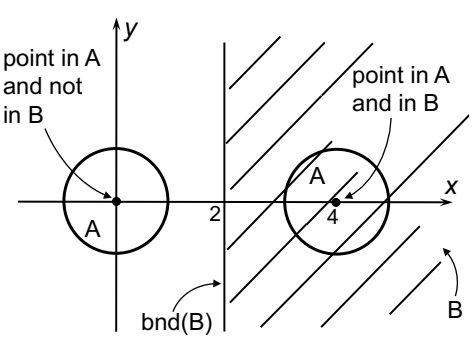


The key to what will ultimately be the definition of "disconnected" is to ask if we already have in our repertoire of definitions anything which is rather "curve-like". The answer is that we have: boundaries tend, in a very rough and general way, to be "curve-like things which separate the interior of a set from the exterior of that set". [Recall, for example, the situation for the disk.] Thus, we do not really need to find fancy definitions of "curve", "one side of the curve" and "the other side of the curve"'. We can use as surrogates "the boundary of a set", "the set itself" (which one may think of as being "on one side of the boundary"), and "the points outside of that set" (which one thinks of as being "on the other side of the boundary"). These observations, then, suggest that we try to capture the essential intuitive content of "there exists a curve such that part of the set is on one side, part on the other, and none of the set meets the curve" as follows.

Definition. Set A in the plane will be said to be *disconnected* if there exists a set B in the plane such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B).

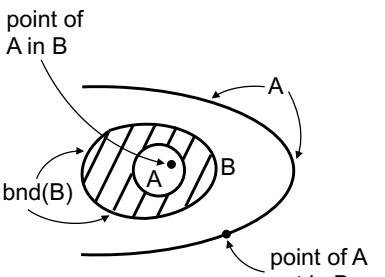
Here, if I have ever seen one, is an example of a good idea in mathematics. One combines the intuition we have of "boundary", the fact that we have formulated a definition of "boundary", and what we want to say by "disconnected" to obtain a candidate for a precise definition of "disconnected". To come up with original ideas, such as this one, for a definition is an activity to which many mathematicians devote a great deal of effort. [This definition is an old one. It is not, of course, original with me.] Of course, we do not wet know whether or not this definition will turn out to be an appropriate one, i.e., whether it will give the expected answer in simple examples, but it certainly looks promising, What remains, then, is to simply try it out.

Let us return, then, to our example of two disks. Is this set, according to our definition, disconnected? Can we find a set B such that some point of A is in B, some point is not in B, and no point of A is in bnd (B)? It turns out that we can. [To find such a set B, one thinks about how we arrived at the definition. One, in one's mind's eye, draws "a curve between the two pieces of A". Then one takes "one

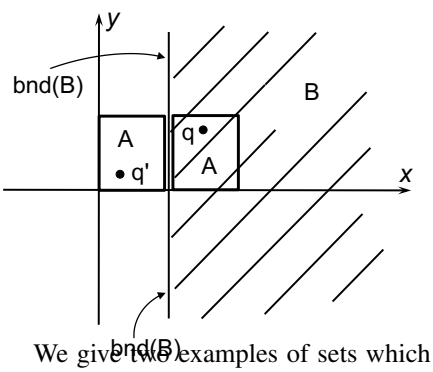


side of that curve", and designates that set
"B".] Let, then, B be the set of all points (x, y) with $x > 2$ – the set illustrated in the figure. Then certainly some point of A is in B (namely, for example, the center of the "right" disk the point (4, 0)) and some point of A is not in B (namely, "left" disk, the point (0, 0)). Further, bnd B is just the line indicated, i.e., the set of all points (x, y) with $x = 2$. [One, of course, must check with the definition of "boundary" to get this.] Clearly, no point of bnd (B) is in A. So, this A satisfies the conditions of the definition (for we have found a set B such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B). So, this set A is disconnected. Similarly, a set A consisting of two points is disconnected.

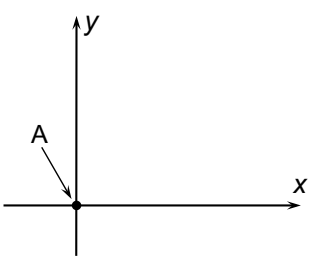
Next, consider the set A consisting of the "horseshoe and disk", as on page 14. This set is also disconnected. Let B, for example, be the set shown. [Again, we will not bother to write down equations.] Then, again, some point of A is in B, some point of A is not in B, and no point of A is in bnd (B). So, this set is disconnected.



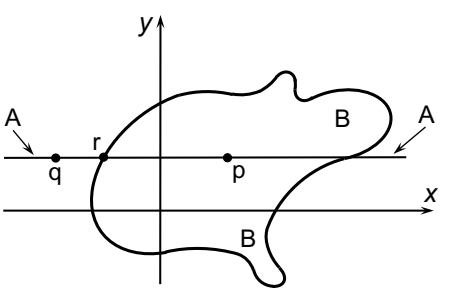
Our "two squares" set A, as illustrated on page 9, is also disconnected. Now, let B consist of all points (x, y) with $x > 1$. Then bnd (B) is all (x, y) with $x = 1$. Now, there is a point (such as q in the figure) in A and in B, a point (such as q') in A and not in B, while no point of A is in bnd (B). [For the latter, the reason is that the "line between the two squares" through which "the boundary of B passes" is not included in our set A.] So, this B will do the job. This set A is disconnected.



We give two examples of sets which are connected (= not disconnected). Let A be the set consisting of the single point at the origin. Does there exist a set B such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B)? Hardly, for A only has one point, and this point would have to be either in B or not in B, no matter what B is. There could hardly be *both* a point of A in B and a point of A not in B, since A has only one point.



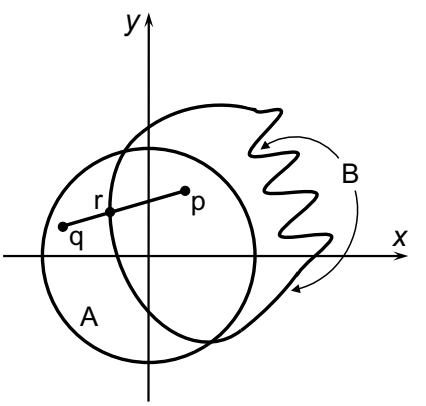
Now consider, as a second example, a line, e.g., the set A consisting of all (x, y) with $y = 1$. This set, we claim, is connected. [A full proof of this assertion, while not terribly difficult, does involve some technical properties of the real numbers. Since we are here not particularly interested in the real numbers, we shall just give an informal argument.] Suppose, on the contrary, that one could find some set B such that some point



of the line A is in B, some point of A is not in B, and no point of A is in bnd (B) (i.e., a set B which could cause A to be disconnected). We argue for a contradiction. Let p be the point of A in B, as shown in the figure. Then p cannot be in bnd (B) (for no point point of A is to be in bnd (B)). That is, it must be false that, for every positive ϵ, some point within ϵ of p is in B and some within ϵ of p is outside of B. But certainly for every positive ϵ, some point within ϵ of p is in B, namely p itself (clearly within ϵ of p and clearly in B). So, the "other part" must fail. That is, it must be false that, for every positive ϵ, some point within ϵ of p is outside of B. That is, there must be some positive ϵ such that *every* point within ϵ of p is in B. But this just means that the point p is in int (B). Next, let q be the point of A not in B, as shown in the figure. Then this q is certainly *not* in int (B), for q is not even in the set B itself. Now start at the point p (which is in int (B)) and work toward point q (which is not in int (B)), and continually ask, as one moves along, whether one is still in int (B). Eventually, the answer will have to change from "yes" to "no". Let r be the point at which this change takes place. [The existence of such an r requires the "technical property of the real numbers".] This r, we now claim, must be in bnd (B). Indeed, for every positive ϵ, there is a point within ϵ of r and in B (since the points "just to right of r along the line" are of course in B). Furthermore, for every positive ϵ, there is a point within ϵ of r and not in B (for should this fail, i.e., were it the case that for some positive ϵ, every point within ϵ of r is in B, then r would be in int (B). But r was chosen to be the first point which ceases to be in int (B).) Putting these two together, we conclude that r is in bnd (B). Now we have our contradiction, for the point r is certainly in A (i.e., on our line) and is also in bnd (B), while our B was supposed to have the property that no point of A is in bnd (B).

The above is, admittedly (and unfortunately), a rather complicated argument, but I hope that its sense is clear. In order that the line A be disconnected, one must find set B such that the line gets from a point (p) in B to a point (q) not in B, all without encountering any point of bnd (B). But this cannot happen, because a transition must somewhere (on the line) take place, yielding a point of the line in bnd (B).

In a similar way, one argues that the disk is disconnected. Suppose that one could find a set B such that some point of A is in B, some point of A is not in B, and no point of A is in bnd B, as shown. The line segment joining p (in A and in B) and q (in A and not in B) will lie entirely within our disk A. But this line segment goes from a point p in B to a point q not in B, and so, arguing as before, we find a point r of this line segment in bnd (B). We thus obtain our contradiction, for this r is in A (since r is on our line segment, and the whole line segment is in A), and also in bnd (B).



Similarly the entire plane and the empty set are both connected.

As an example of all four definitions, we consider again the set A of the second figure on page 2: the set of all (x, y) with x rational. In this case, bnd (A) will be the entire plane (for no matter what point p and what positive ϵ one chooses, there will be some point within ϵ of p and in A, and also some point within ϵ of p and not in A. Int (A) will be the empty set (for, for no point p can one find a positive ϵ such that every point within ϵ of p is in our set A). This A will not be bounded (for, for no point p and positive number c will every point of A be within c of p), and it will be disconnected (for, choosing, say, B the set of all points (x, y) with $x > \pi$, there will certainly be a point of A in B and also a point of A not in B, while no point of A is in bnd (B), for bnd (B) is the set of all points (x, y) with $x = \pi$, while, since π is not rational, none of these points will be in A).

We shall make extensive use of these, as well as other, similar, examples. The skills which, it seems to me, are important here are the ability to apply the definitions mechan-

ically in unfamiliar situation (which is not easy, because the definitions are filled with "for some positive ϵ", "every point of", "for some point", "and also", and so on), and the ability to recognize quickly what the answer is in more familiar situations.

4. Properties of Boundary, Interior, Bounded, and Connected

We began with certain intuitive notions about sets in the plane. By introducing the usual distance between points in the plane, we were able to "capture" these notions, i.e., to formulate mathematical definitions of "boundary", "interior", bounded" and "connected'. The next step was to test these definitions and understand what they are saying. In particular, one wants to know whether or not the definitions faithfully reproduce our own ideas about what "boundary", "interior", "bounded" and "connected" should mean. This "testing and understanding" is normal operating procedure following the introduction of any definition in mathematics. The procedure usually consists of two stages. In the first stage, one applies the definition to various examples (preferably, to examples as diverse as possible), in order to get a feeling for the scope of the definition and how it works. We have now completed this stage. In the second stage, one looks around for properties of the definition: results, usually of an elementary sort, which involve the thing defined in some way, i.e., which relate it to itself or to other mathematical ideas with which one might be familiar. [The line between these two stages is in practice not as sharp as suggested here.] We now turn to this second stage. We emphasize at the outset that the properties of "boundary", "interior", "bounded", and "connected" number well into the hundreds, and probably into the thousands. It is thus not practical to try to list *all* of the properties: nor is it practical to try to memorize those properties one does find. Rather, one wishes to obtain an overall feeling for what is true and what is false, what argument is likely to yield a proof of an assertion and what argument is not, what examples are likely to show that an assertion is false and what examples are not.

Is there any simple relation between a set A and its interior? Let us try a few examples. The interior of the entire plane is the entire plane; the interior of the disk is the disk; the interior of a line is the empty set; the interior of a point is the empty set; the interior of the empty set is the empty set. Can one see any pattern to these examples? Thinking about these for a while, and possibly trying a few more, one notices that there is indeed a pattern: the interior of a set seems never to be "larger" than the original set. So, we have now discovered what seems to be a pattern in various examples. We next wish to see if we can formulate some precise statement which reflects the pattern we have found. [The statement above is unsuitable, for we have no defined "larger".] In this example, it is easy to say what one really means, namely that int (A) never "gets outside" of A, i.e., that every point of int(A) is also in A. Finally, we face the task of deciding whether this pattern is just a coincidence or a special property of our particular examples, or whether it will hold for *all* examples. What does it means for a point p to be in int (A)? It means that, for some positive ϵ, *every* point within distance ϵ of p is in A. But what about p itself? It is certainly within ϵ of p (for, in fact, $d(p,\ p) = 0$). So, it would have to be in A too. Clearly, then, it will be true for *any* example (i.e., for any set A) that every point of int (A) is in A. So, we have found a pattern, expressed it, and decided that it will actually hold for any example. The final step is to express all this in the "theorem-proof" format of mathematics. For the present case, a suitable rendition is:

Theorem. Let A be a set in the plane. Then every point of int (A) is in A.

Proof: Let point p be in int (A). Then, for some positive number ϵ, every point within distance ϵ of p is in A. But p is within distance ϵ of p, and so p must be in A.

The writing of theorems and proofs in mathematics is very much an art form. As such, it is a skill which is learned more through exposure to numerous examples than trough following rules. We shall, of course, see many examples of theorems and proofs. A few general guidelines at this point may, however, be useful. Consider first the words

and symbols which follow "theorem". These should be full sentences (in which symbols, of course, are allowed). These words must represent a clear and unambiguous assertion that something is true. Thus, for example, "Let A be a set in the plane." would not be appropriate, for it does not assert anything, while "Let A be a disconnected set in the plane. Then A consists of two or more pieces." would not be appropriate either, for, while it certainly seems to be trying to assert something, it is not clear exactly what (for we do not know what "consists of two or more pieces" means).

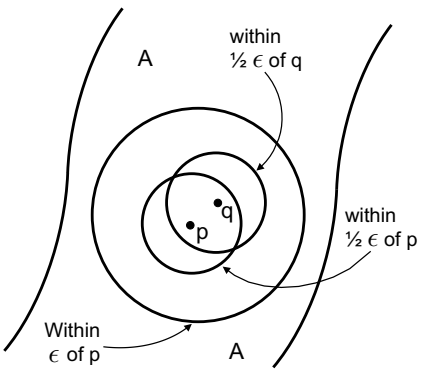
One does not ordinarily omit even obvious conditions in the statement of a theorem. Thus, for the above, "Every point of int (A) is in A.": What is A? a point? a number? [One is in practice a bit more flexible than this. Completely obvious conditions are sometimes omitted.] It is, however, always incorrect to omit significant conditions. If, for example, the theorem above were true only when the set A is bounded, then the omission of this condition (even if it were made clear in the proof that A has to be bounded, and even if the preceding discussion was all about bounded sets) would be inappropriate. Finally, there should be nothing in the statement of the theorem which is irrelevant to the assertion it makes. The following, for example, would not be good theorems: "Let A and B be sets in the plane. Then every point of int (A) is in A."; "Let A be a set in the plane. Then every point of int (A), which represents all points which are 'well inside' A, is in A." A proof should, of course, consist of sentences, and represent a clear demonstration that the assertion of the theorem is in fact true. The proof should be the most direct line one can find, and should of course contain nothing irrelevant to what is being demonstrated. Thus, for example, replacing the first sentence by "Last point p be in int (A), so p is 'well inside' A", replacing the last sentence by "What about the point p itself? This p is within distance ϵ of p and so p must be in A." or adding a sentence at the end such as "For, were p not in A, then it could not be within ϵ of p, and that would give a contradiction, since p is within 0 of p, and so is within ϵ of p." would not be appropriate. Every sentence of the proof should be clear: It should be clear what one is using, what one is concluding, and why. Thus, "Since for some positive every point is within ϵ of p is in A, p must be in A". is just not clear. [What is this p? Are we assuming that p is in int (A)?] The various steps in the proof should be approximately equal in difficulty. That is, instead of five very simple steps followed by one very difficult one, one would combine some of the five initial steps and break up the last one. The difficulties of the successive steps should be such that one can see clearly how to get from one to the next. The following proof, for example, contains too many steps. "To show that every point of int (A) is in A, it suffices to select one point, p, in int (A), assuming nothing more about p, and show that this p is in A. But since p is assumed to be in int (A), it must satisfy the definition.

That is, for some positive ϵ, every point within ϵ of p must be in A. But, no matter what the positive ϵ is, the point p itself is within ϵ of p, for in fact the distance of p from p is zero. So, since every point within ϵ of p must be in A, and since p is within ϵ of p, the point p must be in A. So, since we began with any point p in int (A), and showed that it is in A, we have shown that every point in int (A) is in A." Finally, one does not normally let any motivation or other thoughts one might have creep into the proof.

Is there any simple relation between set A and its boundary? Is, for example, every point of bnd (A) in A, or every point of A in bnd(A)? Let A be the disk (all (x, y) with $x^2 + y^2 < 1$), so bnd (A) consists of all (x, y) with $x^2 + y^2 = 1$. In this example, there are points of A not in bnd (A), and points in bnd (A) not in A. This example, however, suggests that no point can be in both bnd (A) and A. Is this true? Unfortunately, it is not in general: For A a line, bnd $(A) = A$, and so A and its boundary have points in common. There is, apparently, no simple relationship between A and its boundary.

Since the boundary or interior of a set is again a set, one can take the boundary or interior again. Are there any simple relations? Consider int (int (A)), i.e., the set obtained by first determining the interior of A, then the interior of the resulting set. We try some examples. For A the entire plane, int (A) is the entire plane, and so int (int (A)) is the entire plane; for A the disk, int (A) is the same disk, and so int (int (A)) is the disk; for A

a line, or a point, or the empty set, int(A) is the empty set, and so int (int (A)) is also the empty set.



After perhaps trying a few more examples, one realizes that there is a pattern here: What one gets at the second step seems always to be just what one gets at the end. We are led, then, to guess that, for any set A, int (int (A)) = int (A). Is this true in general? Suppose, first, that we have a point of int (A)). Then, by the Theorem on page 18 applied to "int (A)", it follows that our point must be in int (A). So, we have only to show the reverse. The idea is the following. Suppose p is a point of int (A). Then all points within some ϵ of p are in A, as shown in the figure. Now consider the disk centered at p, with radius $1/2\,\epsilon$. All of the points in this disk are of course in A. But all of these points must also be in int (A), for clearly, since all points of the larger disk are in A, all points of the smaller disk must be in int (A). The points of A "fill" the larger disk, and so A includes the smaller disk in its interior. We see, then, that our guess is in fact true. All this might be expressed as follows.

Theorem. For any set A in the plane, int (int (A)) = int (A).

Proof: By the Theorem on page 18, every point of int (int (A)) is in int (A). For the converse, let p be any point of int (A). Then for some positive ϵ, every point within ϵ of p is in A. Let q be any point within $1/2\,\epsilon$ of p. Then every point within $1/2\,\epsilon$ of q must be within ϵ of p, and so must be in A. So, q must be in int (A) Since every point within $1/2\,\epsilon$ of p is in int (A), p is in int (A), p is in int (int (A)).

Are there other such relations? What about int (bnd (A))? For the entire plane, bnd (A) is the empty set, so int (bnd (A)) is the empty set: for A the disk, bnd (A) is a circle, so int (bnd (A) is the empty set; for A a line, or a point, or the empty set, bnd (A) is just A, so int (bnd (A)) is the empty set. The pattern is obvious: We always get the empty set. So, one might guess that always int (bnd (A)) is the empty set. One might at this point try for a while to see if a proof could be found – but one would fail. In fact, this is false. Consider, for instance, the example at the bottom of page 16: A is all (x, y) with x rational. Then bnd (A) is the entire plane, and so int (bnd (A)) is the entire plane! There is no simpler expression for int (bnd (A)).

For S any set in the plane, the *complement* of A, denoted A^C, is the set of all points not in A. [Note that, given the unambiguous statement of what is in and what is not in A, we obtain an unambiguous statement of what is in and what is not in A^C,] For example, the complement of the entire plane is the empty set; the complement of the disk is the set of all (x, y) with $x^2 + y^2 \geq 1$ (the plane with "a hole in it"). The complement of the complement of A is again the set A.

What relations are there between the boundary of interior of A^C and A? Consider int (A^C). For A the disk, A^C is the set of all (x, y) with $x^2 + y^2 > 1$, and so int (A^C) is the set of all (x, y) with $x^2 + y^2 > 1$. For A consisting of the single point $(0, 0)$, A^C consists of all points except $(0, 0)$, and so int (A^C) = A^C. For A the entire plane, A^C is the empty set, and so int (A^C) is the empty set. There is in fact a pattern here, namely that no point is in both A and int (A^C). This is true in general. Indeed, were a point in both A and int (A^C), then, by the Theorem on page 18, it would be in both A and A^C, which contradicts what "complement" means.

A more interesting relation involves the boundary. Imagine taking a photograph of the set A, with A itself in white and the background (A^C) in black. Then the negative of this photograph (with A^C in white and A in black) would be the photograph of A^C. Now think of bnd (A) as the "border" between the black and white regions of the original photograph. Then one might expect that this would be the same as the border between

the white and black regions of the negative. These intuitive remarks suggest:
Theorem. Let A be a set in the plane. Then bnd (A^C) = bnd (A).
Proof: Point p is in bnd (A) if and only if, for every positive number ϵ, there is some point within ϵ of p and in A, and also some point within ϵ of p and not in A. But, since the points in A are precisely those not in A^C, this holds if and only if, for every positive number ϵ, there is some point within ϵ of p and not in A^C, and also some point within ϵ of p and in A^C, But this holds, finally, if and only if the point p is in bnd (A^C).

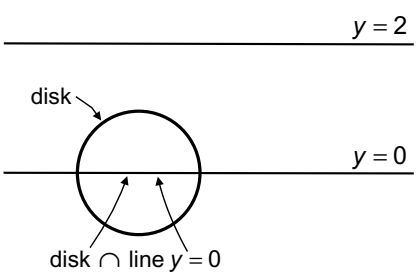
[In mathematics, "... if and only if ···" means "whenever ··· holds, ... also holds, and furthermore whenever ... hods, ··· also holds."] Our previous two theorems were rather ugly things, designed primarily just to illustrate how one discovers theorems, how one states them, and how one proves them. This theorem, however, is a bit more interesting. Its nice feature is that one guesses it from one's ideas about "boundary". It genuinely seems to reflect an aspect of what a "boundary" should be. It furthermore, with our definition of "boundary" turns out to be true. One's confidence in the appropriateness of the definition is strengthened.

For A and B sets in the plane, we say that A is a *subset* of B, written $A \subset B$, if every point in A is also in B.

For example, every set in the plane is a subset of the entire plane; the set consisting only of (0, 0) is a subset of the disk; the empty set is a subset of any set in the plane. If A is a subset of B and B a subset of C, then A is a subset of C.

There are apparently no simple relations connecting subsets and boundaries. [For example, it is not true that, if $A \subset B$, then bnd $(A) \subset$ bnd (B). Let A consist of the single point (0, 0) and B the usual disk. Then $A \subset B$. But bnd $(A) = A$, and bnd (B) consists of all (x, y) with $x^2 + y^2 = 1$, so it is not true that bnd $(A) \subset$ bnd (B).] For the interior, however, we have
Theorem. Let A and B be sets in the plane, with $A \subset B$. Then int$(A) \subset$ int (B).
Proof: Let p be a point of int (A). Then, for some positive number ϵ, every point within distance ϵ of p is in A. But, since $A \subset B$, every point within distance ϵ of p is in B. Hence, p is in int (B).
One interprets the theorem as saying "the smaller the set, the smaller its interior".

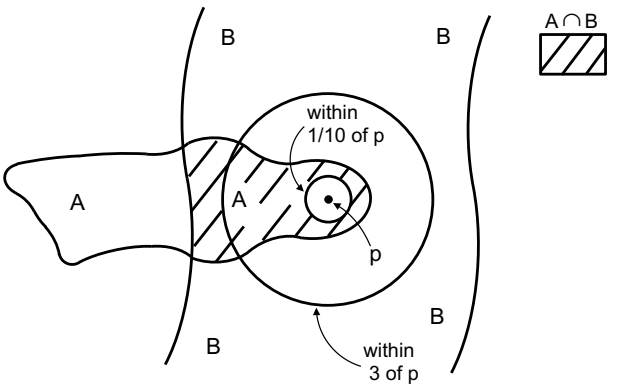
For A and B sets in the plane, the *intersection* of A and B, denoted $A \cap B$, is the set of all points which are in both A and B. Thus, for example, the intersection of the usual disk and the line given by all (x, y) with $y = 0$ is a "segment of a line", given by all (x, y) with $-1 < x < 1$ and $y = 0$; the intersection of this disk and the line given by all (x, y) with $y = 2$ is the empty set. The intersection of any set and the entire plane is that set; the intersection of any set and the empty set is the empty set. For any A and B, $A \cap B$ is a subset of A.



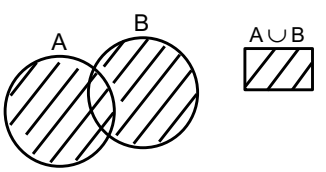
We give one example of a property involving intersections. Think of the points in the interior of a set as those "well inside" the set. Which points would one expect to be "well inside" $A \cap B$? Since $A \cap B$ consists just of those points which are in both A and B, one might expect that the points "well inside" $A \cap B$ will be just those which are both "well inside" A and "well inside" B. One tries a few examples. (For instance, for A the entire plane and B ant set, $A \cap B = B$, and so int $(A \cap B)$ = int (B); but int (A) is the entire plane, so int $(A) \cap$ int (B) = int (B). So, in this example, int $(A \cap B)$ = int $(A) \cap$ int (B).), and finds this to be the case. All this leads to
Theorem. Let A and B be sets in the plane. Then int $(A \cap B)$ = int $(A) \cap$ int (B).
Proof: Let p be a point of int $(A \cap B)$. Then, for some positive number ϵ, all points within ϵ of p are in $A \cap B$. It follows that all points within ϵ of p are in A, whence p is in int (A); and that all points within ϵ of p are in B, whence p in int (B). So, p is in int $(A) \cap$ int (B). For the converse, let q be a point of int $(A) \cap$ int (B). Then q is in int (A), and so,

for some positive number ϵ_1, every point within distance ϵ_1 of q is in A; and q is in int (B), so, for some positive number ϵ_2, every point within distance ϵ_2 of q is in B. Let ϵ, a positive number, be the smaller of ϵ_1 and ϵ_2. Then every point within ϵ of q is both in A and in B, and so is in $A \cap B$. That is, q is in int $(A \cap B)$.

This proof, like all the proofs we have had, consists merely of unravelling the definitions to figure out what one has, and then ravelling then to figure out what one gets. [Indeed, is some sense *all* proofs in mathematics are just this. Theorems in mathematics never have any "real content".] This one, however, is a bit more complicated. for the second part, in particular, the idea is the following. One has point q in int $(A) \cap$ int (B), i.e., q is in both int (A) and int (B), and one wants to show that q is in int $(A \cap B)$. One has positive number ϵ_1, which specifies "how well inside" q is of A, and ϵ_2 for B. What should one choose for one's ϵ to describe "how well inside" q is of $A \cap B$? Suppose, for example, that ϵ_1 were $1/10$, and ϵ_2 were 3, as shown in the figure. That is, q is "just barely well inside" A, but "quite well inside" B. How "well inside" will q be of $A \cap B$? The answer is "just barely well inside", as one sees from the figure. These observations, then, suggest what should be chosen for the ϵ in the proof. [Of course, it is perfectly legitimate to draw a picture to aid in the discovery of a proof. But the picture should not be part of the actual proof.]



For A and B sets in the plane, the *union* of A and B, denoted $A \cup B$, is the set of all points which are either in A or in B. As examples, the union of two disks is the striped region illustrated at the right. The union of any set and the entire plane is the entire plane. The union of any set and the empty set is the original set. For any sets A and B in the plane, A, B, and $A \cap B$ are all subsets of $A \cup B$.



What relations are there between the boundary, interior, and union? From the theorem on the previous page, one might guess that perhaps int $(A \cup B) =$ int $(A) \cup$ int (B). Indeed, this seems somewhat reasonable intuitively, for the points "well inside" $A \cup B$ ought to be those either "well inside" A or "well inside" B. Unfortunately, this is false. Let A be the set of all points (x, y) with x rational, and B the set of all points (x, y) with x irrational (= not rational). Then int (A) and int (B) are both the empty set. Hence, int $(A) \cup$ int (B) is the empty set. But $A \cup B$ is the entire plane (for *every* point (x, y) in the plane has either x rational or x irrational, and so will be in either A or B. In fact, $A^C = B$.) Hence, int $(A \cup B)$ = int (entire plane) = entire plane. So, in this example, we do not have int $(A \cup B) =$ int $(A) \cup$ int (B). In retrospect, we can see what was wrong with our intuition. It is possible, as in this example, to have a point "well inside" $A \cup B$ because A and B work together to "overrun" the vicinity of the point. However, each of A and B individually contains "gaps very near the point", so that the point turns out not to be "well inside", either A or B by itself. Perhaps, however, we can save the other half of our guess. Perhaps every point of int $(A) \cup$ int (B) is in int $(A \cup B)$. This works.

Theorem. Let A and B be sets in the plane. Then int $(A) \cup$ int $(B) \subset$ int $(A \cup B)$.

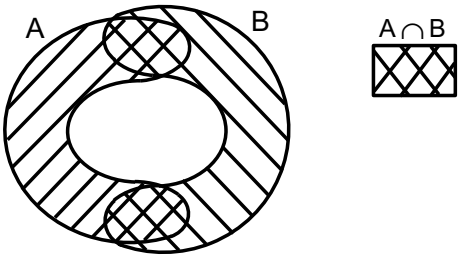
Proof: Let p be a point in int $(A) \cup$ int (B), say p in int (A). Then, for some positive number ϵ, every point within distance ϵ of p is in A. But now every point distance ϵ of p is in $A \cup B$, and so p is in int $(A \cup B)$.

[Note that in this proof we omitted "Since p is in int $(A) \cup$ int (B), p is in either int (A) or

int (B). We shall do the case p in int (A); that for p in int (B) is done similarly."]

We have dealt so far only with properties of the interior and the boundary. We next consider briefly a few properties of bounded and connected.

Is the complement of a bounded set bounded? Certainly not: A disk is bounded but its complement is not. Indeed, quite the opposite is true: The complement of a bounded set can never be bounded. If $A \subset B$ and A is bounded, need B be? No: let A be the disk and B the entire plane. In fact, this is the wrong way around. "Smaller sets" are more likely to be bounded.



Theorem. Let A and B be sets in the plane, with $A \subset B$. Then, if B is bounded, so is A.
Proof: Let B be bounded, and let p be a point and c a positive number such that every point of B is within distance c of p. Then, since $A \subset B$, every point of A is within distance c of p, and so A is bounded.

It follows immediately that the intersection of two bounded sets is bounded (since that intersection is of course a subset of one of the sets). Is the union of two bounded sets bounded?

Is the complement of a connected set connected? No: Let the set be a straight line. Then its compliment is "everything except the line", i.e., two half-planes, and this set is not connected. Is the subset of a connected set connected? The answer is again no. A set consisting of two points is a subset of the entire plane, the entire plane is connected, but the set consisting of two points is not. Is the intersection of two connected sets connected? The answer is again no, but now the example is a bit trickier. The sets illustrated in the figure will do the job. [Since it is clear how one would write the equations for and do in detail (i.e., using the definition) such an example, we will not do so here.] Finally, the union of two connected sets need not be connected: Let each of A and B be a disk of unite radius, but let them be "well separated in the plane". Then each is of course connected, but their union, "two well-separated disks", is not connected.

Does "connected", then, have no simple properties? It turns out that it does have one very nice property. Let us consider again the final example above: the union of two disks. Why, in that example, was $A \cup B$ disconnected? The reason, of course, was that "$A \cup B$ consisted of two pieces, namely A and B separately". It would seem intuitively that this might be the only situation under which A and B are themselves connected while $A \cup B$ is disconnected. Can we make this general idea into a theorem? How do we state "for A and B connected, $A \cup B$ is either connected, or consists just of the two pieces A and B". One possibility, recalling the definition and interpretation of "connected", would be: "Let A and B be connected, and let C be a set such that some point of $A \cup B$ is in C, some point of $A \cup B$ is not in C, and no point of $A \cup B$ is in bnd (C). Then either $A \subset C$ and $B \subset C^C$, or $A \subset C^C$ and $B \subset C$". [This is the right general idea, for, recall, C and C^C are to be the "two pieces" into which C divides the plane.] In fact, this statement is true. It turns out, however, that there is a simpler way to state the same idea. We simply *demand* that A and B be not "completely separate from each other" (so they cannot be the "two pieces" of $A \cup B$) by demanding that they have a point in common.

Theorem. Let A and B be connected sets in the plane, and let p be a point of both A and B. Then $A \cup B$ is connected.
Proof: Suppose, for contradiction, that C were a set such that some point of $A \cup B$ (say, u in A) is in C, some point of $A \cup B$ (say, v in B) is not in C, and no point of $A \cup B$ is in bnd (C). Then no point of A and no point of B is in bnd (C). But point u in A is in C and no point of A is in bnd (C), and so, since A is connected, every point of A must be in C. In particular, point p must be in C. But now point p in B is in C, point v in B is not in C, and no point of B is in bnd (C), which contradicts the fact that B is connected.

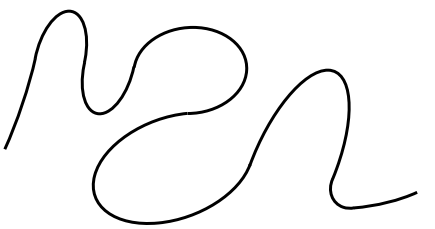
I hope that the structure of the proof is clear. We suppose that $A \cup B$ could be "separated into two pieces" by means of set C. In order for this to be a "genuine separation", some point (u in A) must be in C, and some point (v in B) must be out of C. But A itself cannot be "separated into two pieces" by C, because A is assumed connected. So, A must be either all in C or all out of C. But A has the point u in common with C, and so A must end up "all in" C. But now, having dragged all of A into C, we have also brought the point p (common to both A and B) into C. Now C goes to work on B. Since a little bit of B (namely, the point p) has been dragged into C, and since B is also connected, all of B must be in C too. Here, then, is a contradiction, since the point v of B was supposed to be not in C. [Of course, one could put u and v in other places in $A \cup B$, with the same result.]

This is a nice theorem. It expresses very well an intuitive connotation of "connected", and it is even true.

We emphasize that the purpose of this section has been to illustrate what one does to "understand" a definition in mathematics. The process above is what a mathematician would do automatically on being confronted with a new definition. Of course, the various theorems and examples above are only to illustrate a few of the many possibilities: They are not to be memorized. A second purpose was to illustrate the type of thought-process one goes through in inventing theorems and inventing proofs. Finally, the stated theorems and proofs are intended to be examples of that "art form".

5. Continuous Curves

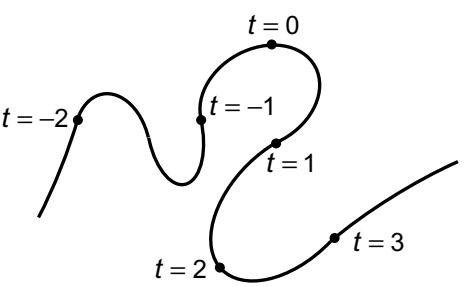
We conclude our discussion of the plane with a final set of definitions. These are perhaps a bit more complicated than those we have treated in the previous sections. Here, we shall merely motivate the definitions, state them, and indicate what they mean. We shall return to this subject later.



One has an intuitive idea of what a "curve" in the plane ought to be: It is a "line", or "what results if one presses a piece of a chalk against a blackboard, and then moves the chalk about". [In mathematics, "curve" is always a generic term, which does not exclude "straightness". Thus, a straight line is also to be a "curve".] Our first goal is to formulate a definition of "curve".

Let us begin by looking for a definition of the form "*A* set *A* in the plane is a curve provided ... " The "..." is supposed to express the idea that the set *A* must be "one-dimensional". How shall we say this idea? Recall that, for *A* a straight line ("one-dimensional"), int (*A*) is the empty set, while for *A* a disk ("not one-dimensional") int *A* is not the empty set. Thus, one might be tempted to replace "..." by "int (*A*) is the empty set". This, however, is not a very good choice. For example, let *A* be the set of all points (x, y) with x rational. Then int (*A*) is indeed the empty set, but we would certainly not wish to call this *A* a "curve". Alternatively, recalling that boundaries tend to be "one-dimensional", one might replace "..." by "for some set *B*, *A* = bnd (*B*)". But this is not too good either, for, for example, the entire plane is the boundary of *some* set *B* (namely, *B* all (x, y) with x rational), while we would certainly not wish to regard the entire plane as a "curve". A few more failures along these lines convince one that this is not a very good task. [Of course, one possibility at this point would be that there is not going to be any "good" definition of a curve. But in this case, as it turns out, there will be.]

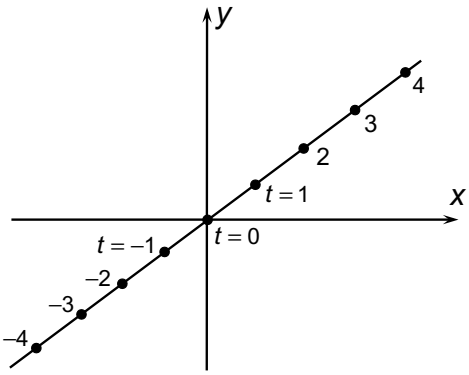
The key to getting a definition of "curve" is to think again about the chalk on the blackboard. In the attempt above, we were trying to characterize a curve in terms of just the set *A* of points of the plane "visited by the chalk". But intuitively a curve would seem to be much more than merely "the set of all points visited".



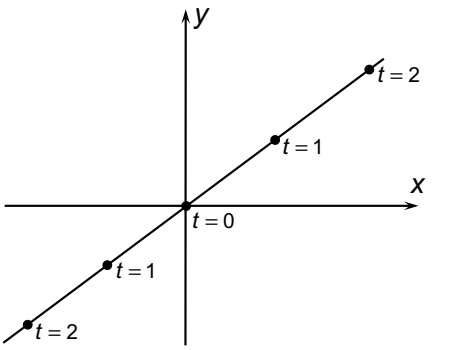
There is also involved, for example, the order in which theses points are visited as the chalk moves over the blackboard. Can we somehow find a definition which somehow brings to mind the actual physical process of moving chalk over blackboard? We can. Imagine that a clock were running as the curve is being drawn on the blackboard. Then at any time t, as read by the clock, the chalk will be at some position on the blackboard. At various other times, the chalk will be at various other positions on the blackboard. Thus, as "time" runs through its various values, "position of the chalk" will run through various points of our plane. The idea, then, would be to regard a "curve" as, not merely the set of points of the plane "visited", but also as the "labelling" of those points by a parameter t, the "time". These ideas suggest.

Definition. A *curve* in the plane is a rule which assigns unambiguously, to each real number t, a point of the plane.

We shall represent a curve symbolically by a letter such as γ. Then, for γ a curve and t a real number, we denote by $\gamma(t)$ that point of the plane assigned to the number t by the unambiguous rule which is the curve. Of course, one thinks of this t as the "time" as recorded by the clock, and of the point $\gamma(t)$ of the plane as the "location of the chalk on the blackboard at time t".

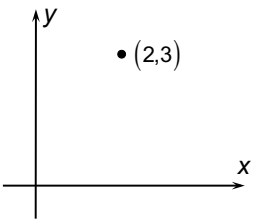


We give some examples of curves. Let γ be the rule which assigns, to the real number t, the point of the plane with $x = t$ and $y = t$, i.e., the point with coordinates (t, t). This is certainly an unambiguous rule: For example, it assigns to the number 93 the points of the plane with coordinates $(93, 93)$. This curve is illustrated in the second figure at the right.

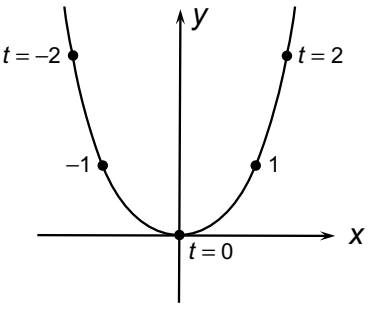


It corresponds physically to "moving the chalk at uniform speed along a straight line tilted 45^0 from the vertical". Next, consider the rule which assigns to number t the point $(2t, 2t)$ of the plane. Again, this is an unambiguous rule: To the number -7, for example, it assigns the point $(-14, -14)$ of the plane. This curve is that shown. Note that this curve describes the same set in the plane as our previous curve. It is, however, to be regarded as a different curve, because we have a different rule. [To the number $t = 3$, for example, the first curve assigns the point $(3, 3)$; the second, the point $(6, 6)$.] Physically, we are now "moving the chalk twice as fast along this line".

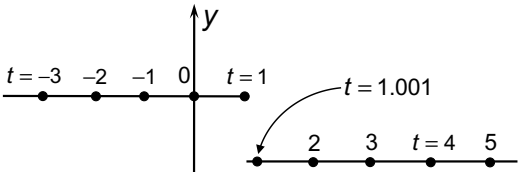
Let γ be the rule which assigns, to number t, the point $(2, 3)$ of the plane. This is unambiguous: It assigns, to $t = -137.9$, the point $(2, 3)$. This is thus a curve. it just "stay put" at the point $(2, 3)$. Consider the parabola $y = x^2$. This is *not* a curve according to our definition. It is not an unambiguous assignment of a point of the plane to each number t – and in fact it does not even say anything about t. What point does this "curve" assign to $t = 5$?



However, consider the rule which assigns, to number t, to point (t, t^2). [For example, to $t = 7$ is assigned the point $(7, 49)$. Clearly, this is "really" $y = x^2$.] Now, we do have a curve. In short, one does not specify a curve by giving "y as a function of x"; rather, one must give "each of x and y as a function of t". The rule which assigns to number t the point $(t^{13} - \cos t + t/(1 + t^2),\ t^6 + t \sin t)$ of the plane is a curve. The rule which assigns to the number t the point of the plane which is $(2, 3)$ if t is a rational number, and $(9t, 61)$ if t is an irrational number is a curve. All we demand for a curve is an unambiguous rule (which, of course, assigns a unique point, $\gamma(t)$, of the plane to each number t) – nothing more and nothing less.



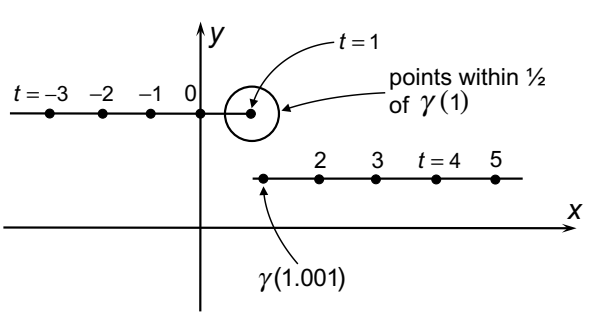
Some curves are better than others. Compare, for example, the curves illustrated in



the figures on pages 25 and 26 with the curve of the last example above, or with the curve given by: For $t \leq 1$, let $\gamma(t)$ be the point $(t, 2)$, and for $t > 1$ let $\gamma(t)$ be the point $(t, 1)$ of the plane. This curve is illustrated on the right. [Note that the above is indeed an unambiguous rule to get from t's to points of the plane, hence, a curve. But "for $t \leq 1$, let $\gamma(t)$ be $(t, 2)$, and for $t \geq 1$ let $\gamma(t)$ be $(t, 1)$" is not a curve, for it is ambiguous. It assigns *two* points of the plane, namely $(1, 2)$ and $(1, 1)$, to $t = 1$.] The physical difference between the curves illustrated on pages 25 and 26 and the curve illustrated above is that the former can be "drawn without ever lifting the chalk from the blackboard", while the latter cannot. Our next task is to capture, by means of a definition, this distinction between curves which do and do not "skip".

To see what will be involved in making this definition, let us return to the curve (with a "skip") illustrated above. At what time did the skip take place? At $t = 1$. Clearly, we must isolate something unpleasant about what the curve was doing around $t = 1$. Where was the chalk at $t = 1$? At the point $\gamma(1)$, namely, the point $(1, 2)$, of the plane. What is it about the behaviour of the curve around $t = 1$ that leads us to regard the curve as having a "skip" there? The behaviour is the following. Whereas for $t = 1$ the chalk is at $\gamma(1) = (1, 2)$, for t just a tiny bit greater than 1, say $1 = 1.001$, the chalk is way over at $(1.001, 1)$. The problem, then, is that, during this "very short time" (from $t = 1$) to $t = 1.001$), the chalk has moved a "rather large amount" in the plane, namely from $\gamma(1) = (1, 2)$ to $\gamma(1.001) = (1.001, 1)$. Thus, to say that there is a skip at $t = 1$, we want to say that "when t moves just a little bit away from the value $t = 1$, the corresponding point, $\gamma(t)$ of the curve moves more than a little bit away from the point, $\gamma(1)$, of the curve corresponding to $t = 1$". All we have to do, then, is express these "little bits" precisely to obtain our definition. In the above example, the "amount of the skip", the "distance $\gamma(t)$ jumped between $t = 1$ and $t = 1.001$", was just a touch more than one unit. Thus, a statement which is true for this curve, and which reflects the fact that it has a skip at $t = 1$, is the following: There are $t-$values as close as you wish to $t = 1$ such that $d(\gamma(t), \gamma(0)) \geq 1/2$. [How close do you want the $t-$values to be to $t = 1$?. Within .01 of 1? Then $t = 1.001$ is this close. But $\gamma(1.001) = (1.001, 1)$, while $\gamma(1) = (1, 2)$, and so $d(\gamma(t), \gamma(1))$ is slightly more than one, and is certainly greater than or equal to $1/2$. Do you want the $t-$values still closer to 1?. You say you want them within.0000001 of $t = 1$? Then $t = 1.00000001$ is this close. But $\gamma(1.00000001, 1) = (1.00000001, 1)$, and its distance from $\gamma(1) = (1, 2)$ is still lightly more than one, and so is certainly greater than or equal to $1/2$.] A more precise version of "as close as you wish" in our statement above yields the following: Given any positive number δ, there is a $t-$value within δ of 1 (i.e., with $1 - \delta \leq t \leq 1 + \delta$) such that $d(\gamma(t), \gamma(1)) \geq 1/2$. The positive number δ represents of course "how close you wish t to be to 1." Instead of "t amount δ of 1", or "$1 - \delta \leq t \leq 1 + \delta$", we can write $|t - 1| \leq \delta$. Recall that $|\quad|$ is the absolute value, the operation of reversing the sign if the number enclosed is negative and keeping the sign if it is positive, so, for example, $|6| = 6$, $|19| = 19$, while $|-3| = 3$ and $|-31| = 31$. So, $|t - 1| \leq \delta$ just says that "the number t is within amount δ of 1, i.e., is between $1 - \delta$ and $1 + \delta$. In short, the "distance" between numbers a and b is just $|a - b|$.

So far, then, we have the following statement, which reflects the fact that there is a skip in the curve illustrated on the previous page: Given any positive number δ, there is a number t satisfying $|t - 1| \leq \delta$ and such that $d(\gamma(t), \gamma(1)) \geq$



1/2. A statement that there is
"no skip" at $t = 1$ would thus
be: There exists a positive number δ such that, whenever number t satisfies $|t - 1| \le \delta$, $d(\gamma(t), \gamma(1)) \le 1/2$. What this says is that " there exists *some* range of $t-$values around $t = 1$ (namely, the range of $t-$values consisting of those t between $1 - \delta$ and $1 + \delta$) such that, for all t in this range, the corresponding point $\gamma(t)$ of the plane is within distance $1/2$ of the point $\gamma(0)$". For example, consider again our curve at the end of page 26. Is it true that there is a range of $t-$values around $t = 1$ such that, for every such t, we have $d(\gamma(t), \gamma(1)) \le 1/2$? It is certainly not true. The locus of points within distance $1/2$ of $\gamma(1)$ is the disk shown in the figure. there is no range of $t-$values around $t = 1$ such that for every t in this range $\gamma(t)$ lies within this disk. No matter what "range" (i.e., positive number δ) is given, one can find a $t-$value within this range (namely, a $t-$value very slightly greater than 1) such that $\gamma(t)$ is outside of this disk.

We now have practically the statement we are after. But where did this "$1/2$" come from? It arose originally from our example. We were looking at a curve with a skip of about 1, so we choose $1/2$ as the "amount of skip" to give ourselves a little leeway. What the statement above really says, then, is that there is no skip at $t = 1$ of "amount" greater than $1/2$. But we are not just interested in skips of amount greater than $1/2$: We want to rule out skips of *any* amount. We do so by: For *every* positive number ϵ, there exists a positive number δ such that, whenever t satisfies $|t - 1| \le \delta, d(\gamma(t), \gamma(1)) \le \epsilon$. This is to hold for *every* positive ϵ (the"amount of the skip"). It is to hold for $\epsilon = 1/2$ (no skips of amount greater than $1/2$), for $\epsilon = 1/10$ (no skips of amount greater than $1/10$), and so on. It thus says that there are no skips of any amount at $t = 1$. Finally, the discussion above has been for skips at the $t-$value $t = 1$. One can proceed similarly for other $t-$values. We thus arrive at:

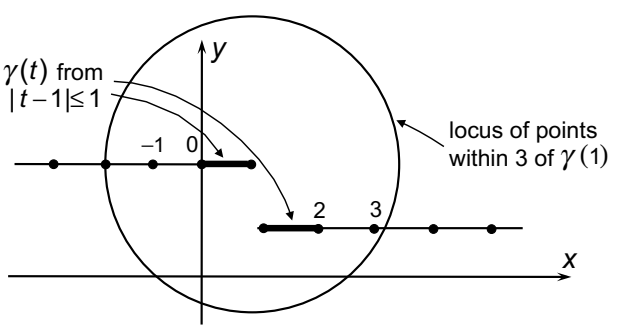
Definition. Curve γ is said to be *continuous* at $t = t_0$ if, for every positive number ϵ, there exists a positive number δ such that, whenever $|t - t_0| \le \delta$, $d(\gamma(t), \gamma(t_0)) \le \epsilon$.

Thus, continuous at $t = t_0$ means "having no skips of any amount at $t = t_0$". For "having no skips of any amount anywhere", we have

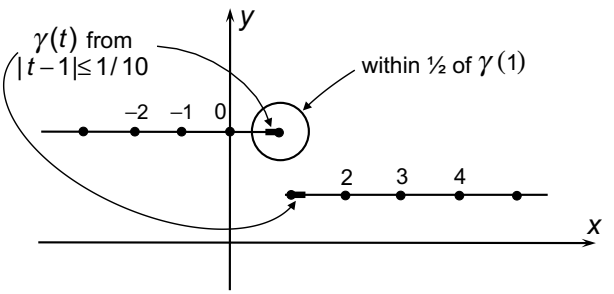
Definition. Curve γ is said to be *continuous* if, for every number t_0, γ is continuous at $t = t_0$.

As usual in mathematics, one invests all one's thoughts and ideas in the *formulation* of a definition. However, once the definition has been stated, one essentially ignores all of one's ideas, and just mechanically follows the rules as laid down by that definition. Thus, in order to show that a curve is continuous at $t = t_0$, one has to figure out how, given any positive number ϵ, one can find a positive number δ such that, whenever $|t - t_0| \le \delta, d(\gamma(t), \gamma(t_0)) \le \epsilon$. We give some examples.

Consider the curve illustrated on the previous page. Is it continuous at $t = 1$? Is it true that, for every positive number ϵ, there exists a positive number δ such that, whenever $|t - 1| \le \delta$, $d(\gamma(t), \gamma(1)) \le \epsilon$? Let's try an ϵ, say $\epsilon = 3$. Is it true that there exists a positive number δ such that, whenever $|t - 1| \le \delta, \gamma(t)$ is within 3 of $\gamma(1)$? The locus of points within 3 of $\gamma(1)$ is the disk shown. Does there exist a positive number δ such that, whenever $|t - 1| \le \delta, \gamma(t)$ is in this disk? There certainly does: Choose $\delta = 1$. Then $|t - 1| \le 1$ means that t is between 0 and 2, and so the corresponding $\gamma(t)$'s are those shown in the figure. But this part of the curve is certainly within our disk. So, there does exist a positive number δ such that, whenever $|t - 1| \le \delta, d(\gamma(t), \gamma(1)) \le 3$. So, we tried $\epsilon = 3$, and it worked. But we have to

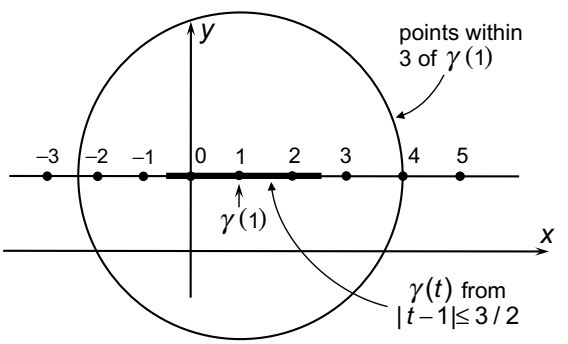


show this for *every* positive ϵ.

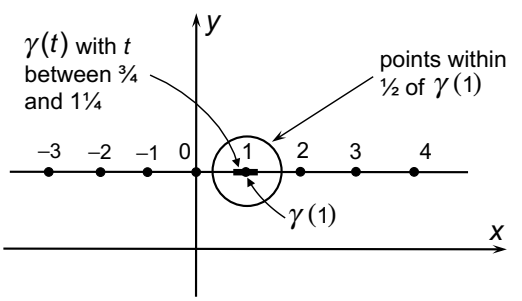


Let's try another, say $\epsilon = 1/2$. Is it true that there exists a positive number δ such that, whenever $|t-1| \leq \delta, \gamma(t)$ is within $1/2$ of $\gamma(1)$? The locus of points within $1/2$ of $\gamma(1)$ is the disk shown. Does there exist a positive number δ such that, whenever $|t-1| \leq \delta, \gamma(t)$ is in this disk? There does not. What value for δ will work? What about, say, $\delta = 1/10$. Is it true that, whenever t is between $9/10$ and $1\,1/10$, $\gamma(t)$ is in this disk? It is not true, as one sees from the figure. Clearly, *no* δ will work. Thus, there does not exist a positive δ such that, whenever $|t-1| \leq \delta, d(\gamma(t), \gamma(1)) \leq 1/2$. So, it is not true that, for *every* positive ϵ, there exists a positive δ such that, whenever $|t-1| \leq \delta, d(\gamma(t), \gamma(1)) \leq \epsilon$ [for, as we have just seen, there is none for $\epsilon = 1/2$]. So, our curve is not continuous at $t = 1$. [This example gives the clue as to what one should choose for ϵ in order to show that a given curve is not continuous. Choose ϵ to be smaller than the "amount of skip".]

Is that curve continuous? Is it true that, for *every* number t_0, γ is continuous at $t = t_0$? It is not true, for this curve is not continuous at $t = 1$. Our definitions thus give the expected answers for this example.



Consider, as a second example, the curve γ with $\gamma(t) = (t, 1)$. is this curve continuous at $t = 1$? Is it true that, for every positive ϵ, there exists a positive δ such that, whenever t is within δ of 1, $\gamma(t)$ is within ϵ of $\gamma(1)$? Let's try $\epsilon = 3$. Does there exist a positive δ such that, whenever t is within δ of 1, $\gamma(t)$ is within 3 of $\gamma(1)$? The locus of points within $3 of \gamma(1)$ is the disk shown. We want to find a positive δ such that, whenever t is within δ of 1, $\gamma(t)$ is within 3 of $\gamma(1)$. But there is such a δ, namely for example. $\delta = 3/2$. The portion of the curve for $|t-1| \leq 3/2$ is indicated in the figure, and all this portion is within our disk. So, we found a δ for this ϵ.



What about another, say $\epsilon = 1/2$? Does there exist a positive ϵ such that, whenever t is within δ of 1, $\gamma(t)$ is within $1/2$ of $\gamma(1)$? The locus of points within $1/2$ of $\gamma(1)$ is the disk shown. What do we choose for δ? Well, $\delta = 1/4$ will do, for, whenever t is within $1/4$ of 1 (i.e., between $3/4$ and $1\,1/4$), $\gamma(t)$ is within $1/2$ of $\gamma(1)$, as shown in the figure. Thus, we found a suitable δ for this choice of ϵ. One tries another, say $\epsilon = 1/10$. For this ϵ, one will have to choose a still smaller δ, for determines "how close" t will be to 1, and we are going g to have to make t "quite close" to 1 if we are going to guarantee that $\gamma(t)$ will be within only $1/10$ of $\gamma(1)$. A suitable choice is $\delta = 1/20$. It is indeed true that, whenever t is within $1/20$ of 1, $\gamma(t)$ is within $1/10$ of $\gamma(1)$. We thus conclude that it is true that for every positive ϵ, there exists a positive δ such that whenever $|t-1| \leq \delta, d(\gamma(t), \gamma(1)) \leq \epsilon$ [namely, given positive ϵ, we choose $\delta = 1/2\epsilon$]. [We have not of course actually proven

formally that this δ always works, but it is clear, and a full proof is not difficult.] We thus conclude that our curve is continuous at $t = 1$.

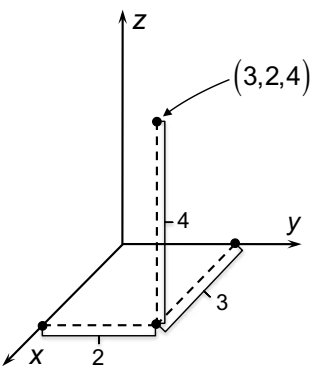
Is this curve continuous at $t = 5$? Yes, by exactly the same argument. Is it continuous at every $t-$value? Yes. So, this curve, since it is continuous at every $t = t_0$, is continuous.

Similarly, one verifies that the three curves illustrated on page 25 are all continuous. The curve γ with $\gamma(t) = (1, 1)$ if t is rational and $\gamma(t) = (4, 5)$ if t is irrational is not continuous at any $t = t_0$. [The curve "jumps between these two points every time t changes from a rational to an irrational and back". To show not continuous, one must find an ϵ such that the definition will not work. Since here the "skip", the distance between these points, is 5, one would choose for ϵ any number less than 5.]

6. Metric Spaces

So far, we have introduced four intuitive notions involving sets in the plane (or, if "continues curve" is included, five), expressed those intuitive notions as mathematical definitions, given various examples of those definitions, and proven a number of theorems which relate the definitions to each other. We are now at a crossroads: What shall we do next? One possibility would be to go ahead and look for more intuitive notions, more definitions and more theorems. It turns out, however that there is a more interesting and fruitful direction. It consists essentially of trying to isolate what is really relevant to what we already have, and trying to generalize it. It is the direction of "quality" as opposed to quantity. In this section, we begin this program.

Why have we been dealing, all this time, with the plane, as opposed to some other space? The reason, frankly, is that it is easy, for the plane, to draw figures to illustrated what is going on. Suppose, however, that one got interested in Euclidean three-dimensional space (the space of solid geometry)? A point of this space would be labeled by three numbers, so the coordinates of a typical point would be (x, y, z). One could, of course, introduce the notion of a set in this Euclidean three-dimensional space (an unambiguous statement of which (x, y, z) are in the set and which are out.). One would furthermore have similar intuitive ideas (of boundary, interior, bounded, and connected) about such sets. How would one go about capturing such intuitive ideas as definitions? The first step would be to observe that one also has a notion of geometrical distance in Euclidean three-dimensional space. Indeed, for $p = (x, y, z)$ and $q = (x', y', z')$ two points in this space, the distance between these points is given, by the Pythagorean theorem just as before, by



$$\mathrm{d}(p, q) = \sqrt{(x' - x)^2 + (y' - y)^2 + (z' - z)^2}$$

So, one might decide, one will use this distance to formulate definitions of "boundary", "interior", "bounded", and "connected". One expects that finding suitable definitions this time will be somewhat easier, since we already have the experience gained with the plane.

But, were one to embark on such a program, a strange thing would happen. Not only does finding suitable new definitions turn out to be "somewhat easier" than it was before, it further turns out that one can just repeat, word for word, the old definitions, replacing "point of the plane" and "set in the plane" everywhere by "point of Euclidean three-dimensional space" and "set in Euclidean three-dimensional space". Thus, for example, the definition of "interior" on page 11 would become: "Let A be a set in Euclidean three-dimensional space. Then the *interior* of A is the set of all points p [now, of Euclidean three-dimensional space] such that, for some positive number ϵ, every point q [now, of Euclidean three-dimensional space] with $d(p, q) \leq epsilon$ is in A." The resulting four definitions not only make sense (i.e., are genuine mathematical definition), but furthermore accurately reflect the intuitive ideas they are to capture. One would thus instantly obtain all four definitions.

One would next seek theorems which relate these definitions. But now a second miracle takes place, One would soon realize that all our old theorems remain true for Euclidean three-dimensional space (with, again, the mere replacement of "plane" everywhere by "Euclidean three-dimensional space"). Not only this, but the old proofs are again proofs of the new theorems (with the usual replacement). Thus, for example, we would translate the theorem on page 18 as follows:

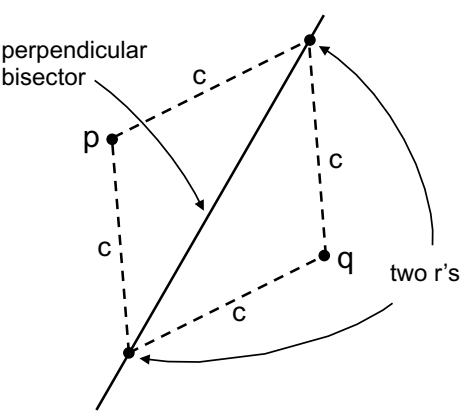
Theorem. Let A be a set in Euclidean three-dimensional space. Then every point of int (A) is in A.
Proof: Let point p in int (A). Then, for some positive number ϵ, every point within distance ϵ of p, is in A. But p is within distance ϵ of p and so p must be in A.
[Now, of course, "p" is a point of Euclidean three-dimensional space.]

In short, we have essentially already done this subject for Euclidean three-dimensional space.

So, one might turn to another possibility, say, one-dimensional space. The "coordinate" of a point would just be one number, (x), so our space is a single "line". Again, one can introduce sets in this one-dimensional space: again, one can introduce the geometrical distance between points (for $p = (x)$ and $q = (x')$ points in our one-dimensional space, set $d(p, q) = |x - x'|$, the geometrical distance between the points along the line); again, we can just repeat our old definitions (replacing "plane" by "line") to reflect our intuitive ideas of "boundary", etc.; again the old theorems remain theorems and the old proofs remain proofs. We have already done this subject for the line, too.

What would be one's response to this state of affairs? Purchase a large quantity of paper, and proceed to rewrite everything we have done two more times, but with this minor changes? Just say "similarly for Euclidean three-dimensional space, and for the line", we have above? The response of the mathematician would be the following. These "miracles" are an indication that we did not fully understand what we were doing in the first place. We were *talking* plane, but *doing* much more. A fuller understanding of what is really going on here is to come by somehow isolating what is genuinely relevant to our four notions and the relationships between them. What we want to do, then, is to generalize – to find a broader framework for our discussion which will at one stroke and encompass the plane, Euclidean three-dimensional space, the line, and hopefully even more possibilities.

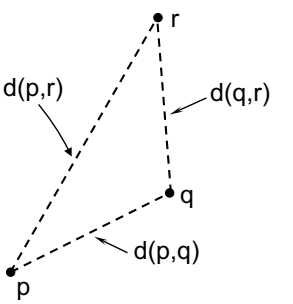
What is the crucial thing which made our treatment of sets in the plane possible? What is the essential ingredient which our treatments of Euclidean three-dimensional space, the plane, and the line have in common? This is obvious: It is this distance d. Now comes the hard part. We have to figure out what it was about our distances that made things work out as they did for the plane: what the "distances" for Euclidean three-dimensional space, the plane, and the line have in common; what features our distances have which reflect our intuitive, geometrical ideas of what "distance" should be like. It is a problem not unlike that of inventing, for the first time, a definition of "boundary". It is, in the fullest sense, a problem in mathematics.



An example of an inappropriate property of distance is the following. For any two distinct points p and q, and any number c greater than $d(p, q)$, there are exactly two points r with $d(p, r) = c$ and $d(q, r) = c$. This property is true for the plane, Indeed, the two points lie on the perpendicular bisector of the line segment joining p and q, as shown in the figure. But this property does not hold is Euclidean three-dimensional space 9in which there is an infinite number of such "r' s "), or in the line (in which there is no such "r"). Furthermore, we never used this property, and it does not at all seem to be mandated by our intuitive idea of "distance". It is much more a property of "plane" than of "distance".

We now obtain the "distance-like" properties of our distances. The key to the first property is the theorem on page 18 (int $(A) \subset A$). What property of distance was used there? It was used in the part of the proof which reads "But p is within distance ϵ of p,

. . .". Of course, the reason why p is within distance ϵ of p is that $d(p, p) = 0$. Here, then, is a property of distance (that $d(p, p) = 0$) which is true for Euclidean three-dimensional space, etc. which was used in our treatment of the plane, and which seems natural for distance. [If I have to accept "the distance of p from p is five units", then I do not want to think about that "distance".] A closely related property, which has been implicit in our discussion of the plane, is: If $p = q$ then $d(p, q)$ is positive. [That is, one cannot have distinct points with zero distance between them, or two points distance – 7 apart.] The second property is such a simple and natural one that, while it has always been in the background, it has never been mentioned explicitly. [These are often the most difficult properties to find!] It is that, for any two points p and q, $d(p, q) = d(q, p)$. The distance from p to q is the same as the distance from q to p. Again, this property would seem to be an integral part of what we would wish to mean by "distance".



The third property was used, for example, in the theorem on page 20, in the sentence "Then every point within $1/2\epsilon$ of q must be within ϵ of p, . . ." in the proof. Let r be such a point (within $1/2\epsilon$ of q). Then what is being used here is the following; If $d(p, q) \leq 1/2\epsilon$ and $d(q, r) \leq 1/2\epsilon$. then $d(p, r) \leq \epsilon$. Why is this true? Think of a triangle, with vertices p, q and r. Then if the distance from p to q is no more than $1/2\epsilon$ and the distance from q to r is no more than $1/2\epsilon$, the distance from p to r could hardly be more than $\epsilon (= 1/2\epsilon + 1/2\epsilon)$. It could hardly be that going from p to r via q is actually shorter than going "directly" from p to r. That is, it could hardly be that $d(p, q) + d(q, r)$ (distance from p to r via q) is less than $d(p, r)$ (distance from p to r directly). The property which is being used, then, is: $d(p, q) + d(q, r) \geq d(p, r)$. This is called the *triangle inequality*. It is true for the plane, for the line, and for Euclidean three-dimensional space. It was used in one of our theorems. Again, it seems like a natural property for "distance".

It turns out that an appropriate list of properties to reflect "distance" is just the three given above. In practice, of course, one would have to try many other properties, trying to make the decision, for each, whether or not it should be included. One would presumably also have second thoughts about these three. In any case, one eventually settles on these three properties. Having decided on the properties, one organizes everything as follows.

Definition. A *metric space* is a set X together with a rule which assigns, to any two points p and q of X, a number, $d(p, q)$, such that the following conditions are satisfied;

1. For any points p and q of X, $d(p, q)$ is zero if $p = q$ and positive if $p \neq q$.
2. For any points p and q of X, $d(p, q) = d(q, p)$.
3. For any points p, q and r of X, $d(p, q) + d(q, r) \geq d(p, r)$.

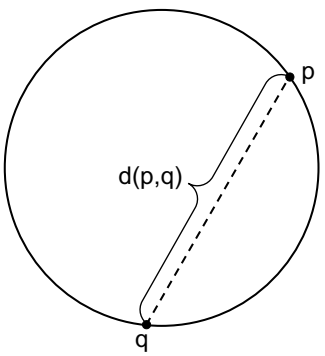
A metric space, then, is to represent "the idea of distance" in its purest form, with no other extraneous features". A metric space consists of a set X (the "points between which distance will be given"), and the rule which assigns (unambiguously, of course) a number $d(p, q)$ to any two points p and q (which is to be regarded as, and will be called, the *distance* between p and q). But all these assignments of distances cannot be made just by whim, at least if one expects to obtain a metric space. They must be made in such a way that these three conditions are satisfied, such that they have at least some of the character of a "distance". Of course, to specify a metric space, one must say what the set X is, one must give some unambiguous rule for obtaining a number, $d(p, q)$, for *any* two points p and q of X, and one must have our three conditions satisfied. Nothing more and nothing less. In particular, it is not necessary that the metric space be "reasonable" in any other sense. [After all, we had the option, prior to the definition, to demand "reasonable", and in fact we took advantage of that option to impose our three conditions. Now, it is too late to ask for "still more reasonable". Of course, if one happened to come up with some nice additional conditions, nothing would stop one from defining a new thing to incorporate them.]

We now give some examples of metric spaces.

Let X be the plane, i.e., the set of all pairs, (x, y) of real numbers (certainly, a set). For p and q any two points of the plane, let $d(p, q)$ be the number given by the formula on page 6 (certainly, an unambiguous assignment of a number to any two points of X). Then our three conditions are of course satisfied, as we have already seen. This, then, is a metric space.

Let X be Euclidean three-dimensional space, i.e., the set of all triples, (x, y, z), of real numbers. For p and q any two points of X, let $d(p, q)$ be the number given by the formula on page 31. Our three conditions are certainly satisfied. This is a metric space. Similarly, the line (with $d(p, q)$ for $p = (x)$, $q = (x')$, given by $|x - x'|$) is a metric space.

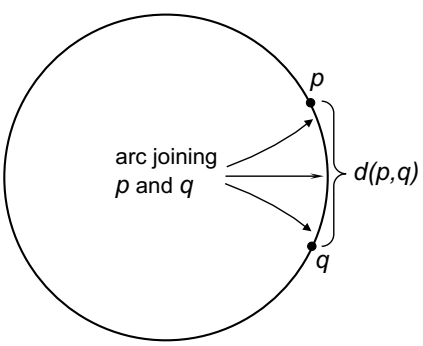
Let X be the circle, i.e., the set of all pairs, (x, y), of numbers, with $x^2 + y^2 = 1$. Let, for p and q any two points of X, $d(p, q)$ be the geometrical distance between these points on the plane (i.e., use again the formula on page 6). Then the first condition is certainly satisfied ($d(p, q)$ is zero if $p = q$, and is positive otherwise); and the second condition id certainly satisfied (always, $d(p, q) = d(q, p)$).



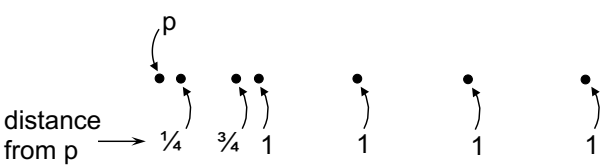
Finally, the third condition is also satisfied. Indeed, this is immediate, since points of X are "really" points of the plane (although, of course, not all points of the plane get to be in X), and the triangle inequality is satisfied for the plane.

Let X be the same circle as on this page, and let, for p and q any two points of X, $d(p, q)$ be the length of the shortest segment of arc of the circle connecting p and q. Thus, for p and q as shown, $d(p, q)$ would be the length of the arc shown. [Note that this is an unambiguous assignment which yields a number in exchange for two points.] Again, the first two conditions are immediate. One furthermore easily convinces oneself that the third condition is also satisfied. We have a metric space. [Which one is right, this metric space or the previous one? What is the real distance between points on the circle? There is no "right" or "real" in mathematics. To give a metric space, one must give a set X and the distances. That is it. The definition does not require that anything be "real".]

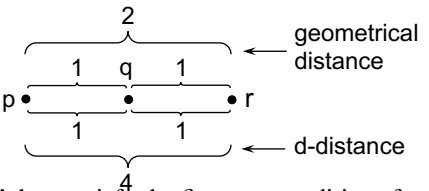
Let X be the plane. For p and q any two points of X, let $d(p, q)$ be the number which is zero if $p = q$, and one if $p \neq q$. [This is an unambiguous, if "unreasonable", rule. Thus, $(0, 0)$ and $(.0001, 0)$ are distance 1 apart: $(0, 0)$ and $(23,926, -189)$ are distance 1 apart.] Let's see if the conditions are satisfied. For $p = q$, $d(p, q)$ is certainly zero. For $p \neq q$, $d(p, q)$ is one which is certainly positive. The first condition is ok. furthermore, $d(p, q) = d(q, p)$ (since both sides are zero if $p = q$ and both sides are one if $p \neq q$). Finally, we come to the third condition. How could this possibly be violated? Since all "distances" are either one or zero, the only possibility would be if $d(p, q)$ is zero, $d(q, r)$ is zero, while $d(p, r)$ is one. But this could never happen, for $d(p, q) = 0$ would mean $p = q$, and $d(q, r) = 0$ would mean $q = r$ – and so we would have to have $p = r$, and so $d(p, r) = 0$ also. We conclude, then, that even the triangle inequality is satisfied. This is a metric space!



Let X be the plane. For p and q any two points of X, let $d(p, q)$ be the usual geometrical distance between p and q if that distance is less

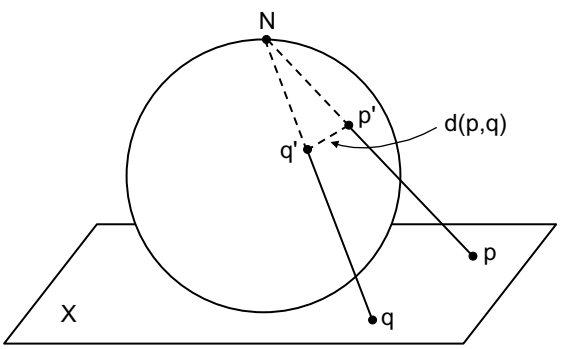


than one, and one if that geometrical distance is greater than or equal to one. The distances (with this d) of some points from p are shown in the figure. Arguing as in the previous example, one sees that this is a metric space. [Again, the only tricky step is the triangle inequality. One must consider several cases, according to whether the geometrical distances between p, q and r are less than or greater than one.]



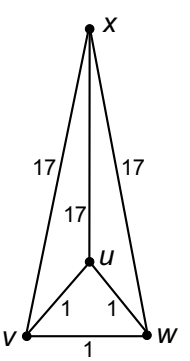
Let X be the plane. Let, for p and q any two points of X, $d(p, q)$ be five times the geometrical distance between p and q. This is a metric space.

Let X be the plane. Let for p and q any two points of X, $d(p, q)$ be the square of the geometrical distance from p to q. This "distance" does satisfy the first two conditions for a metric space. However, it does not satisfy the third condition. Let, for example, p, q and r be as shown (so the geometrical distance from p to q, and from q to r, are both one). Then we would have $d(p, q) = 1^2 = 1$, $d(q, r) = 1^2 = 1$, but $d(p, r) = 2^2 = 4$. It is not true that $d(p, q) + d(q, r) \geq d(p, r)$. This is not a metric space.



Let X be the plane. Place a unit sphere on X, as shown in the figure, and let N denote the "North pole". For p and q any two points of the plane, we obtain a number $d(p, q)$ as follows. First draw a straight line from p to N, and identify the point p' at which that line enters the sphere. Similarly, identity the point q'. Now take the straight-line distance (in our three-dimensional space of the figure) between p' and q'. This distance is to be the number $d(p, q)$. Note that this is indeed a rule to get a number from any two points of X. Points "very far out near the edge of the plane", e.g., (23, 234, 9.001), yield points p' "very near the North pole N". Thus, any two such points will, from our d, be "very close together", even though they appear geometrically to be very far apart. That the first two conditions are satisfied is immediate. It actually turns out (although it is perhaps a bit complicated to check in detail) that the triangle inequality is satisfied. This is a metric space.

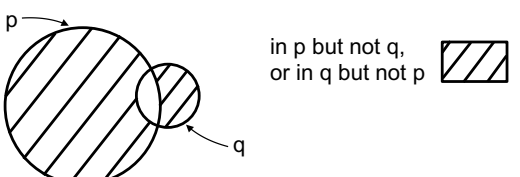
Let X be the set consisting of the one point p, and set $d(p, p) = 0$. This is a set and a rule. The three conditions are easily checked. This is a metric space. Let X be the set consisting of just two points, p and q. Set $d(p, p) = 0$, $d(q, q) = 0$, $d(p, q) = 19$. This is a metric space.



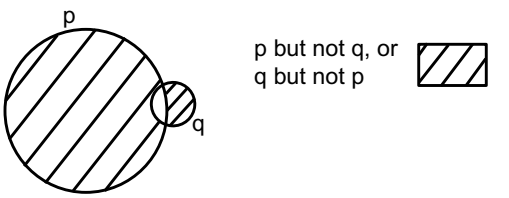
Let X consist of just four points, with distances as indicated in the figure, and with $d(p, p)$ always zero. There are a total of sixty-four possibilities for p, q and r in the triangle inequality. One checks them. [It is actually easy: Almost all the cases reduce to just a couple.] The first two conditions are immediate. This is a metric space. If one of the "17's" above were changed to a "15", one would not have a metric space, for the triangle inequality would be violated. [For which p, q, r?]. If one of the "17's" were changed to "0", the first condition would also be violated, and again we would not have a metric space.

We now consider a couple of more exotic examples.

Let X be the set of all disks in the plane. So, a "point" of X is actually an entire disk. Let p and q be two "points" of X as shown. We obtain

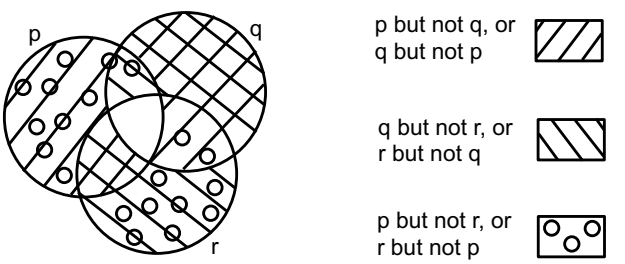


$d(p, q)$ as follows. Consider the set of all points of the plane which are in p (a disk) but not in q, or else are in q but not in p, as in the examples shown above. Then, let $d(p, q)$ be the area, in the plane, of this set of points. Thus, in the first example on the right, $d(p, q)$ is rather large, for the area of the region "in p but not q or in q but not p" is rather large. In the second example, $d(p, q)$ is rather small, i.e., these "points" of the set X are "close together". This X, d we claim, is a metric space. For the first condition, note that $d(p, q)$, as the area of some region in the plane, is always greater than or equal to zero.



Furthermore, in order for $d(p, q) = 0$ there would have to be no points in p but not q, and no points in q but not p, i.e., p and q (points of X) would have to be the same disk in the plane. That $d(p, q) = d(q, p)$, the second condition, is immediate.

Finally, we check the triangle inequality, $d(p, q) + d(q, r) \geq d(p, r)$. How each of the "$d$'s" above of course represents the area of some region in the plane. In order to show this inequality, therefore, it suffices to show that every point of the plane in the region of the plane corresponding to $d(p, r)$ is either in the region corresponding to $d(p, q)$ or in the region corresponding to $d(q, r)$ (or, of course, possibly in both). Having shown this, it will follow that the area one takes to compute $d(p, r)$ cannot exceed the sum of the areas for $d(p, q)$ and $d(q, r)$. Let, then, u be a point in $d(p, r)$, so that u is either in p but not r or in r but not p. We must show that this u is either in the region for $d(p, q)$ or in the region for $d(q, r)$. There are two cases to consider: u in p but not r and u in r but not p. Let us take the first, so u is in p but not r. We now ask: Is u in the set q? If yes, then u (in p but not r, and assumed to be in q) is in q but not r – and so is in the region whose area represents $d(q, r)$. If no, then u (in p but not r, and assumed not to be in q) is in p but not q – and so is in the region whose area represents $d(p, q)$. Similarly for the case u in r but not p. We conclude, then, that this is a metric space. [In terms of the figure above, what we have shown is that "the entire region with circles in it has stripes going one way or the other".]



Similarly, one could let X be the set of all rectangles in the plane, and define d as above, using the area. Or, one could let X consist of disks and rectangles, and again define d using the area. In either case, we obtain a metric space. [Why not just let X be the set of all subsets of the plane, and again define d by using the areas? This will not work, for it would be necessary to be able to say what the "area" of an arbitrary subset of the plane is. What, for example, is the "area" of the set consisting of all (x, y) with x rational?]

Let X be the set of all finite subsets of the positive integers, so a "point" of X is just a finite list of integers, with no repetitions. For example, $p = (19, 3, 197, 11)$ is a point of X. Now consider two such points of X, say this one and $q = (197, 11, 5)$. We compute

$d(p, q)$ as follows. First, strike out any integers which appear in both p and q (in this example, 197 and 11). Then add all the remaining integers (in this example, 19, 3, and 5). The resulting sum is $d(p, q)$, so in this example $d(p, q) = 19+3+5 = 27$. One verifies that this is a metric space. [The argument is essentially the same as for the example above.]

Let X be the set of all ordered pairs of points in the plane, so a point of X, $p = (u, v)$, consists of a first point, u, of the plane and also a second point, v of the plane. [We permit $u = v$.] Given two such pairs, say the one above and $q = (u', v')$, we define the distance between these two points of X by $d(p, q)$ = (geometrical distance between u and u') + (geometrical distance between v and v'). Thus, in order that two points of X be "close together", the corresponding first points of the plane must be close together in the plane, and also the corresponding second points. One easily checks, using for each of the three conditions the corresponding condition for the plane, that this X, d is a metric space.

7. Boundary, Interior, Bounded, and Connected in Metric Spaces

We decided that, in our discussion of properties of sets in the plane, we were talking plane, but actually doing much more. Our attempt to better understand what was going on, to focus on what was really relevant to our discussion, led to the notion of a metric space, "distilled essence of distance". We have some examples of metric spaces – enough examples to see that, while "metric space" does in some sense capture our ideas of distance, it also admits some pretty strange-looking things. The next step is to go ahead and carry out the program for which our metric spaces were invented. We want to repeat our treatment of sets in the plane, but now for "sets in a metric space". We want to carry out this treatment once and for all – for a general metric space. Then the plane, Euclidean three-dimensional space, etc. will all be special cases. In this section, we carry out the first half of this program: We introduce the definitions, with examples, of "boundary", "interior", "bounded", and "connected" in a metric space. One expects, of course, that the new definitions will be essentially the same as the old ones, since it was precisely the fact that our earlier discussion seemed to have a wider domain of applicability which led us in the direction of a metric space in the first place.

We begin with "boundary" and "interior". We have

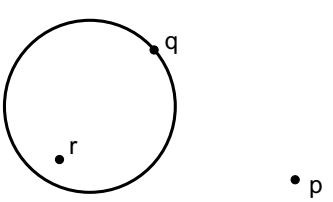
Definition. Let X, d be a metric space, and let $A \subset X$. Then the *boundary* of A is the set of all points p of X such that, for every positive number ϵ, there is a point q of X with $d(p,q) \leq \epsilon$ and q in A, and also a point q' of X with $d(p,q') \leq \epsilon$ and q' not in A.

Definition. Let X, d be a metric space, and let $A \subset X$. Then the *interior* of A is the set of all points p of X such that, for some positive number ϵ, every point q of X with $d(p,q) \leq \epsilon$ is in A.

Note that these are just our earlier definitions, but with a few minor changes. Instead of "A is a set in the plane", we say "A is a subset of X", or $A \subset X$". [In the case in which X *is* the plane, a "set in the plane" is the same thing as a "subset of X".] Further, we include for emphasis the fact that all points considered in the definition are supposed to be points of X. We shall continue to write bnd (A) and int (A) for the boundary and interior of A.

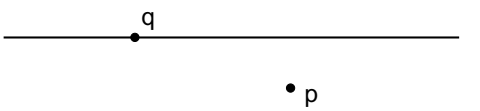
We give some examples.

Let X be the plane, and d the usual geometrical distance. Then these definitions of course reduce to our old definitions of the boundary and interior. Thus, for any set A in this metric space, bnd (A) and int (A) are the boundary and interior we have found before.



Let X be the plane, and let d be: $d(p,q) = 1$ if $p \neq q$, and $d(p,q) = 0$ if $p = q$. This, as we have seen, is a metric space. Let A be the usual disk (all (x,y) with $x^2 + y^2 < 1$). We determine int (A). Consider first the point p (given, say, by $(3,4)$). Is p in int (A)? Is it true that, for some positive ϵ every point of X within distance ϵ of p is in A? It is not true. Indeed, no matter what positive ϵ is chosen the point p itself will be within ϵ of p, while of course p is not in A. So, this p is not in int (A). Similarly, the point q shown in the figure is not in int (A). What about the point r? Is it true that, for some positive ϵ, every point of X within ϵ or r is in A? It is true. Indeed, consider $\epsilon = 1/2$ (certainly, *some* positive ϵ). What is the locus of points within $1/2$ of r? Well, the distance of a point from r is 1 if that point is different from r, and 0 if that point is r itself. Thus, the locus of points within $1/2$ of r is just r itself. [Note; It is *not* a disk of radius $1/2$ centered at r. For "distance", we must of course use the "d" of our metric space.] So, every point within $1/2$ of r (since there

is only one, namely r) is in A. So, r is in int (A). Clearly, then, int $(A) = A$, i.e., we get the same answer as we obtained before. What is bnd (A)? Is point p in bnd (A)? Is it true that, for every positive ϵ, there is a point within ϵ of p and in A, and also a point within ϵ of p and not in A? It is not true. Consider, for example, $\epsilon = 1/2$. The locus of points within $1/2$ of p consists of just p itself. Since p is not in A, there is no point within $1/2$ of p and in A. So, p is not in bnd (A). Is r in bnd (A)? Is it true that, for every positive ϵ, there is a point within ϵ of r and in A, and also a point within ϵ of r and not in A? Again, it is not true: Chose $\epsilon = 1/2$, and so the locus of points within $1/2$ of r consists of r itself, and so there is no point within $1/2$ of r and not in A. Finally, we consider the point q. It is true that, for every positive ϵ, there is a point within ϵ of q and in A, and also a point within ϵ of q and not in A? This is not true either! Try $\epsilon = 1/2$. The locus of points within $1/2$ of q consists of q itself. But q is not in A. So, there is no point within $1/2$ of q and in A. Clearly, then, bnd (A) is the empty set.



Next, let, in this metric space, A be the line, the set of all (x, y) with $y = 1$. Is point p in int (A). Can we find a positive ϵ such that every point within ϵ of p is in A? Certainly not: point p will, no matter what positive ϵ is chosen, be within ϵ of p, while p is not in A. Is point q in int (A)? It is. Choose $\epsilon = 1/2$. Then the locus of points within $1/2$ of q is just q itself. Thus, *every* point $1/2$ of q is in A. Thus, this q is in int (A)! Clearly, int $(A) = A$, i.e., is just the line. What about bnd (A)? Is point p in bnd (A) It is not: For $\epsilon = 1/2$, there is no point within ϵ of p and in A. Similarly, point q is not in bnd (A): For $\epsilon = 1/2$, there is no point within ϵ of q and not in A (since, of course, the locus of points within $1/2$ of q consists of just q itself]. Thus, bnd (A) is the empty set. [In these examples, we used "$1/2$" because it is positive and less than 1. For p a point of X, the locus of points within 2 of p is of course the entire set X.]

We notice a pattern in the examples above. In each case, we had int $(A) = A$ and bnd (A) the empty set. Inspection of the arguments also reveals that the arguments for the two choices of A were essentially the same. In fact, these arguments had virtually nothing to do with A. These observations suggest

Theorem. Let X, d be the metric space with X the plane and $d(p, q) = 1$ if $p \neq q$ and $d(p, q) = 0$ if $p = q$. Thus, for any $A \subset X$, int $(A) = A$ and bnd (A) is the empty set.

Proof: For p in A, every point within $1/2$ of p is in A, and so p is int (A). For p not in A, there could be no positive ϵ such that every point within ϵ of p is in A, for p itself, not in A, will always be within ϵ of p, and so p is not in int (A). So, int $(A) = A$. For p any point of X, the set of points within $1/2$ of p consists of p itself. So, it cannot be the case that for every positive there is a point within ϵ of p and in A, and also a point within ϵ of p and not in A. So, p is not in bnd (A). Hence, bnd (A) id the empty set.

What is going on here? Why do we get answers so different from what we might have expected? Are we to conclude that there is something wrong with the definitions? Our conclusion should be something quite different. What the theorem is trying to tell us is that our intuitive ideas about the plane are intuitively connected with geometrical distance in the plane. Whether we realized it or not, geometrical distance was an integral part of our intuition about "interior" and "boundary". Change the distance (e.g., to obtain the metric space above), and the answers change. It is through the introduction of metric spaces, then, that we come to see the constituents of our ideas about sets in the plane.

Next, let X be the plane, and let d be: $d(p, q)$ is the geometrical distance from p to q if this geometrical distance is less than one, and $d(p, q) = 1$ if that geometrical distance is greater than or equal to 1. This is of course a metric space. The locus of points within distance $1/2$ of point p consists of a disk of radius $1/2$ centered at p; the locus of points within distance $3/4$ of p is a disk of radius of $3/4$ centered at p. But, the locus of points within distance 2 of p is the entire plane (since all "distances" in this metric space are less than or equal to 1). Now let A again be our disk. [See page 38] Is point p in int (A)? It is not. No matter what positive ϵ is chosen, point p will be within distance ϵ or p, while

p is not in A. Is point q in int (A)? Again no, for the same reason. What about point? Is there a positive ϵ such that every point within ϵ of r is in A? There certainly is, say, $\epsilon = 1/10$. The locus of points within $1/10$ of r is the disk of radius $1/10$ centered at r (since $1/10$ is less than one), while all points in this disk are certainly in A. So, r is in int (A). Clearly, int $(A) = A$, i.e., the same answer we got before. What about bnd (A)? Point p is not in bnd (A), for say, for $\epsilon = 1/3$ (so the locus of points within $1/3$ of p is the disk of that radius centered at p), there is no point within $1/3$ of p in A. Point r is not in bnd (A) (for there is no point within $1/10$ of r and not in A). What about point q? Is it true that, for every positive ϵ, there is a point within ϵ of q and in A, and also a point within ϵ of q and not in A? Let us try some ϵ's to see if they work. What about $\epsilon = 5$. Is there a point within 5 of q in A, and also a point within 5 of q and not in A? The locus of points within 5 of q is the entire plane, and so there certainly is. Clearly, there also is for $\epsilon = 10$, or $\epsilon = 2$, or, indeed, for any ϵ greater than 1. What about something smaller than 1, say, $\epsilon = 1/2$. Is there a point within $1/2$ of q and in A, and also a point within $1/2$ of q and not in A? The locus of points within $1/2$ of q is the disk of that radius centered at q. Clearly, there are such points. One now sees that this will work for *every* positiveϵ. Hence, point q is in bnd (A). Clearly, bnd (A) is the "rim" of the disk, i.e. the set of all points (x, y) with $x^2 + y^2 = 1$.

Next, let A be the line as illustrated at the end of page 38 As usual, the point p is not in int (A). Is q in int (A)? We must try to find a positive ϵ such that every point within ϵ of q is in A. Will $\epsilon = 5$ do? No, for the locus of points within 5 of q is the entire plane, and not all these points are in A. Will $\epsilon = 1/2$ do? No, for the locus of points within $1/2$ of q is the disk of that radius centered at q, and not all these points are in A. Clearly, we are not going to find such an ϵ. So, point q is not in int (A) either. Thus, int (A) is the empty set. Is point p in bnd (A)? It is not. For, say, $\epsilon = 1/10$, so the locus of points within $1/10$ of p is the disk of that radius, there will be no point within ϵ of p and in A. Is point q in bnd (A)? Is it true that, for every positive ϵ, there is a point within ϵ of q and in A, and also a point within ϵ of q and not in A? We try a few ϵ's. Consider $\epsilon = 5$, so the locus of points within 5 of q is the entire plane. There certainly is a point within 5 of q and in A, and also a point within 5 of q and not in A. Clearly, $\epsilon = 10$, or $\epsilon = 2$ will work, too. What about, say, $\epsilon = 1/2$? Is there a point within $1/2$ of q and in A, and also a point within $1/2$ of q and not in A? The locus in this case is a disk of radius $1/2$, and certainly there are such points. Clearly, this will work for *any* positive ϵ. So, point q is in bnd (A). We conclude, then, that bnd $(A) = A$, i.e., the same answer as we obtained before for the usual plane.

Again, we see a pattern to these examples. In both cases, we got the same answer as we obtained earlier for the ordinary plane. Looking back at the arguments, one realizes that this is no coincidence. The mechanism is the following. Let us written, as above, d for the distance in the present metric space, and d' for ordinary geometrical distance in the plane. Then, for any points p and q of X, $d(p, q)$ is just $d'(p, q)$ if the latter is less than or equal to one, and is one if $d'(p, q)$ is greater than one. Now fix a subset A of X. Let us write "int" for the interior using distance d, and "int′" for the interior using geometrical distance d'. We want to show that int $(A) = int'(A)$. Consider, then, some point p of int (A). This means of course that there is some positive ϵ such that every q with with $d(p, q) \leq \epsilon$ is in A. We want to show that this p is also in int′ (A). That is we want to find some positive number ϵ' such that every q with $d'(p, q) \leq \epsilon'$ is in A. So, we have ϵ (such that every q with $d(p, q) \leq \epsilon$ is in A), and we are looking for ϵ' (such that every q with $d'(p, q) \leq \epsilon'$ is in A). The question is: What should we choose for ϵ'? To see the answer, recall that $d(p, q)$ is just $d'(p, q)$ if the latter is less than or equal to one, and one otherwise. Clearly, then, we always have $d(p, q) \leq d'(p, q)$, i.e., the effect of our "funny" distance is to regard points as at least as close together as they are geometrically. Because of this, we claim we can just choose $\epsilon' = \epsilon$ and it will work. Does this choice of ϵ' satisfy our condition? Is it true that, every q with $d'(p, q) \leq \epsilon'$ is in A? Yes, for if $d'(p, q) \leq \epsilon'$, then (since $d(p, q) \leq d'(p, q)$), we also have $d(p, q) \leq \epsilon'$, and so (since we

choose $\epsilon' = \epsilon$) we also have $d(p, q) \leq \epsilon$, and so (by the defining property of ϵ) we then have q in A. Thus, we supposed that p was in int (A), i.e., we supposed a positive ϵ such that every q with $d(p, q) \leq \epsilon$ is in A. We then found an ϵ' (namely, ϵ) such that every q with $d'(p, q) \leq \epsilon'$ is in A. That is, we showed that p is also in int$'$ (A). That is, we have shown that every point of int (A) is also in int$'$ (A).

We now do the reverse. We fix a point p in int$'$ (A), and we want to show that this p must be in int (A). That is, we have positive number ϵ' (such that every q with $d'(p, q) \leq \epsilon$ is in A), and we are looking for ϵ (such that every q with $d(p, q) \leq \epsilon$ is in A. What should we choose for this ϵ? What about choosing $\epsilon = \epsilon'$? Let us try an example, say $\epsilon' = 5$. For this example (choosing $\epsilon = \epsilon'$) we would have that every q with $d'(p, q) \leq 5$ is in A, and we wish to conclude that every q with $d'(p, q) \leq 5$ is in A. Does this follow? Well, the locus of points with $d'(p, q) \leq 5$ is a disk of radius 5 centered at p (for d' is just geometrical distance). However, the locus of points q with $d(p, q) \leq 5$ is the entire plane (for, no matter what q is, $d(p, q)$ can be at most one, and so will always be less than or equal to 5). Thus, we would have that every point of the disk of radius 5 about p is in A, and would wish to conclude that every point of the plane is in A. Clearly, this does not follow. Thus, we have ϵ' (such that every q with $d'(p.q) \leq \epsilon'$ is in A, and are looking for ϵ (such that every q with $d(p, q) \leq \epsilon$ is in A. We tried making the choice $\epsilon = \epsilon'$, and found, at least for $\epsilon' = 5$, that it will not work. What ϵ *will* work for $\epsilon' = 5$? We claim that $\epsilon = 1/2$ will work. The locus of points with $d(p, q) \leq 1/2$ is the disk of that radius centered at p, and all these points are within the disk of radius 5 about p. Thus, given that every q with $d'(p, q) \leq 5$ is in A, it follows that every q with $d(p, q) \leq 1/2$ is in A. So, for this ϵ' (namely 5), we have found a suitable ϵ (namely $1/2$). Let us try another ϵ', say $\epsilon' = 137$. For this choice we would have that every q with $d('p, q) \leq 137$ is in A, and we are looking for ϵ such that every q with $d(p, q) \leq \epsilon$ is in A. Again, $\epsilon = 1/2$ will do, for every q with $d(p, q) \leq 1/2$ (i.e., every q in the disk of radius $1/2$ centered at p) will have $d'(p, q) \leq 137$ (i.e., will be in the disk of radius 137 centered at p), and so will be in A. Clearly, then, $\epsilon = 1/2$ will do for these "large" ϵ'. Will $\epsilon = 1/2$ work for *any* positive ϵ'? Let us try $\epsilon' = 1/10$. For this choice, we would have that every q with $d'(p, q) \leq 1/10$ is in A, and we wish to conclude that every q with $d(p, q) \leq 1/2$ is in A. Is this true? The locus of points q with $d'(p, q) \leq 1/10$ is the disk of that radius centered at p: the locus of points q with $d'(p, q) \leq 1/10$ is the disk of that radius centered at p. Can we conclude, knowing that every q in the disk of radius $1/10$ centered at p is in A, that every q in the disk of radius $1/2$ centered at p is in A? We cannot. So, this ϵ (namely, $1/2$) will not work for $\epsilon' = 1/10$. But clearly, an ϵ which *will* work for $\epsilon' = 1/10$ is $\epsilon = 1/10$.

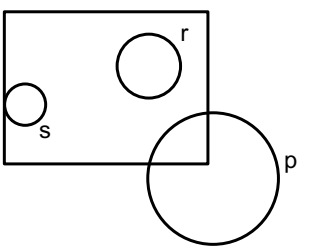
The suitable is now clear. We have ϵ' (such that every q with $d'(p, q) \leq \epsilon'$ is in A), and are looking for ϵ (such that every q with $d(p, q) \leq \epsilon$ is in A). When ϵ' is "large" (specifically, greater than one), we must choose ϵ something like $1/2$ (for if we dare choose ϵ also large, specifically, greater than one, then the locus of points q with $d(p, q) \leq \epsilon$ will become the entire plane, and we shall not be able to conclude that these are all in A). However, when ϵ' is "small" (specifically, less than one), then we must start making ϵ small too (for if we were to continue choosing $\epsilon = 1/2$, then the locus of points q with $d(p, q) \leq \epsilon$ will continue to be the disk of radius $1/2$ centered at p, and this disk will not remain within the locus of points q with $q'(p, q) \leq \epsilon'$). A suitable choice for our ϵ, then, is this: For $\epsilon' \geq 1$, choose $\epsilon = 1/2$; for $\epsilon' < 1$, choose $\epsilon = \epsilon'$. For this choice, it will indeed be the case that, given that every q with $d'(p, q) \leq \epsilon'$ is in A, every q with $d(p, q) \leq \epsilon$ is in A. Thus, we have shown that every point of int$'$ (A) is also in int (A).

Putting our two conclusions together, we conclude that int $(A) = \int'(A)$, for any set A in the plane. The key observations which led us to this conclusion were i) given positive ϵ such that every q with $d(p, q) \leq \epsilon$ is in A, there exists positive ϵ' (namely, $\epsilon' = \epsilon$) such that every q with $d'(p, q) \leq \epsilon'$ is in A, and ii) given positive ϵ' such that every q with $d'(p, q) \leq \epsilon'$ is in A, there exists positive ϵ (namely, $\epsilon = 1/2$ if $\epsilon' \geq 1$ and $\epsilon = \epsilon'$ if $\epsilon' < 1$) such that every q with $d(p, q) \leq \epsilon'$ is in A. By a similar argument, one concludes that bnd (A) (the boundary using distance d) is always the same as bnd$'$ (A) (the boundary using

distance d').

What is one to make of all this? We began with the boundaries and interiors of sets in the plane, using the usual geometrical distance. We now try "changing the distance". For one such change (making the distance zero if the points are the same, and one otherwise), our definitions yield strange answers, which do not at all correspond with our intuition. But for another change (making the distance the geometrical distance if that is less than or equal to one, and one otherwise), our definitions yield precisely the boundary and interior we obtained before. What these examples are doing is giving us detailed information about what facets of "distance" are actually relevant to our intuition of "boundary" and "interior". It is clear that not *everything* about distance counts for the boundary and interior, for at least one significant change in what these distances are to be changes neither the boundary nor the interior. On the other hand, *something* about distance counts, for another change in what the distances are to be changes both the boundary and interior. What, exactly, is it about "distance" which *is* relevant to our ideas of boundary and interior? This is a question to which we shall return shortly.

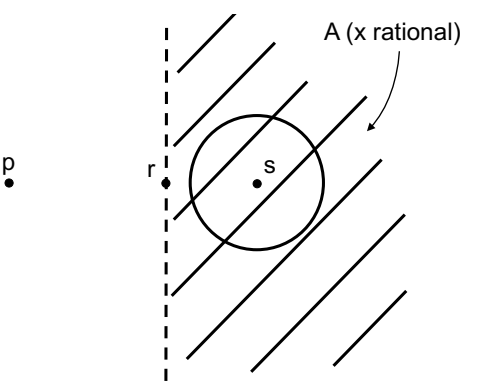
We give a few more examples of interiors and boundaries of sets in metric spaces.



Let X, d be the metric space of page 36, so X is the set of all disks in the plane, and d is computed using areas. Let A be the subset of X consisting of all disks which lie entirely within some fixed square in the plane (say, the square given by all (x, y) with $0 \leq x \leq 1$ and $0 \leq y \leq 1$). Thus, for example, the disks p and s in the figure are not in A. While disk r is in A. [Note that A is indeed a subset of X. We have clearly specified which points of X are in A and which are not.] What is int (A)? The point p of X will certainly not be in int (A): It is not even close to being true that, for some positive ϵ, every point q (disk!) within distance (using areas!) ϵ of p is in A. The point s of X is also not in A. Given positive ϵ, let q be a disk obtained by displacing a very slightly to the left. Then, if this "very slightly" is small enough, q will be within distance ϵ of s. But, of course, this q (no matter how small "very slightly" is) will not be in A. So, it is not true that, for some positive ϵ, every q with $d(q, s) \leq \epsilon$ is in A. So, s is not in int (A). The point r is in int (A). Clearly, then, int (A) will consist of all disks which lie within the square, and whose edge does not touch the boundary of the square. On the other hand, bnd (A) consists of all disks which lie within the square, and whose edge does touch the boundary of the square (for, in order, that disk p be in bnd (A), it must be the case that for every positive ϵ there is a disk within ϵ of p and in A, and also a disk within ϵ of p and not in A).

Let the metric space be the same, but now let A be the subset of X consisting of all disks with center at the origin, $(0, 0)$. What is int (A)? Clearly, the only serious candidates for points in int (A) are disks with center the origin. Is such a disk in int (A)? It is not. No matter what positive ϵ is given, let q be obtained from p (the one with center the origin) by a small displacement, so q will be within ϵ of p, but q will not be in A. So, int (A) is the empty set. On the other hand, bnd $(A) = A$.

Let X be the set consisting of all points (x, y) of the plane with x rational. Let d be the usual geometrical distance. This is a metric space. Let A be the subset of X consisting of all points *of* x with $x > \pi$. Thus, the point $(20, 17)$ is in A, while $(-1, 17)$ is not (because we do not have $x > \pi$), and $(2\pi, 17)$ is not (because this "point" is not even in X: It does not have x rational. What is int (A)? The point p, given by $(-1, 1)$ is not even close to being



in int (A). What about the "point" r, given by $(\pi, 1)$? This is not in int (A), because it is not even in X. It is not even a "point" of the metric space we are considering. Finally, consider the point s, given by $(4, 1)$. Is it true that, for some positive ϵ, every point q of X with $d(p, q) \leq \epsilon$ is in A? This is true! Choose $\epsilon = 1/2$. Then every point q of x with $d(p, q) \leq 1/2$ is indeed in A. [What about the point $q = (1.2\,\pi, 1)$ which would seem to be within $1/2$ of s, and not in A? But this is not a "point" of our metric space X, because its $x-$value is not rational.] Clearly, then, int $(A) = A$. What is bnd (A)? Point p is certainly not in bnd (A). "Point" r is not even under consideration. What about point s? This point s is not in bnd (A), for it is not true that, for every positive ϵ, there is a point within ϵ of s and not in A. Indeed, choose $\epsilon = 1/2$. There is, as we just saw, no point q of X within $1/2$ of s and not in A. So, in this example, bnd (A) is the empty set.

Let the metric space be the same, and let A be the line given by all (x, y) with $x = 1$. Then int (A) is the empty set, and bnd $(A) = A$. Let A be "the closest thing we can have to our disk": All (x, y) with $x^2 + y^2 < 1$ and x rational (a subset of X). Then int $(A) = A$, while bnd (A) consists of all (x, y) with $x^2 + y^2 = 1$ and x rational.

These, then, are some examples of the boundaries and interiors of sets in metric space.

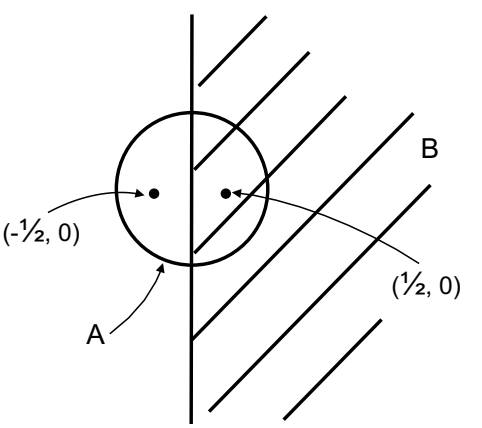
We turn next to the definition of "connected" for a set in a metric space. Again, the appropriate definition is suggested by the corresponding definition (page 14) for a set in the plane.

Definition. Let X, d be a metric space, and let $A \subset X$. Then A is said to be *disconnected* if there exists a subset B of X such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B).

Note that we deal only with the possible disconnectedness of a subset of X. We also require that B be a subset of X. Of course, "bnd (B)" above is the boundary using the metric space X, d. We give some examples.

Let X be the plane, and d ordinary geometrical distance. Then this definition of course reduced to our earlier definition (of disconnected for a set in the plane). Thus, our earlier examples of connected and disconnected sets in the plane apply.

Let X be the plane, and let d be: $d(p, q) = 1$ if $p \neq q$ and $d(p, q) = 0$ if $p = q$ – a metric space. Let, for example, A be the disk: the set of all (x, y) with $x^2 + y^2 < 1$. Is A disconnected? Yes. Let, for example, B be the set of all points (x, y) with $x > 0$, as illustrated on the right. Then some point of A (for example, $(1/2, 0)$) is in A, and some point of A (for example, $(-1/2, 0)$) is not in A. Furthermore, no point of A is in bnd (B), for as we have seen (theorem on page 39) bnd (B) must be the empty set, and so no point whatever is in bnd (B). Thus, A is disconnected. Clearly, the key feature which causes this result is the fact that the boundary of any subset of this metric space is the empty set. In fact, something much more general is true. Let A be any subset of X (in this metric space) having at least two points, say p and q. Then, we claim A must be disconnected. [Choose for B the subset consisting of just the point q. Then some point of A is in B (namely, q), some point of A is not in B (namely, p), and no point of A is in bnd (B) (since bnd (B) is the empty set). Thus, any subset of X having two or more points must be disconnected. On the other hand, a subset of X consisting of just one point, as will as the empty subset of X, are connected.



Nest, let X be the plane, and let $d(p, q)$ = the geometrical distance from p to q if that distance is less than one, and $d(p, q) = 1$ if that geometrical distance is greater than or equal to one. Which subset of X will be disconnected in this metric space? The answer,

as it turns out, is extremely simple. First recall that, for any subset B of X, bnd (B) (for this metric space) is precisely the same set as boundary one would get if the distance were instead the ordinary geometrical distance (what we shall call bnd$'$ (B). It follows from this, we claim, that the disconnected sets in this metric space are exactly the disconnected sets in the ordinary plane. Indeed, suppose that $A \subset X$ is disconnected in this metric space, so there is some $B \subset X$ such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B). Then (since bnd (B) = bnd$'(B)$), it follows that some point of A is in B, some point of A is not in B, and no point of A is in bnd$'$ (B). That is, it follows that A is disconnected in the plane with the usual geometrical distance. For the reverse, suppose that $A \subset X$ is disconnected with the usual geometrical distance, so there is a set $B \subset X$ such that some point of A is in B, some point of A is not in B, and no point of A is in bnd$'$ (B). Then it follows that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B). That is, it follows that A is disconnected also in this metric space. In short, since this "chop the distances' down to one" metric space gives the same boundaries as the ordinary plane, it also gives the same connected and disconnected sets.

Let X be the set consisting of all points (x, y) of the plane with x rational. Let d be the usual geometrical distance. Set $A = X$ (certainly a subset of X). Is A disconnected? The answer is yes. Let $B \subset X$ consist of all points of X with $x > \pi$. Then certainly some point of A is in B (since $A = X$), some point of A is not in B, and yet (since bnd (B) is the empty set) no point of A is in bnd (B). Next, let A be "the closest thing to a disk" in this metric space: all (x, y) with $x^2 + y^2 < 0$ and x rational. Is A disconnected? It is. Let $B \subset X$ consist of all points (x, y) of X with $x > \pi/4$. [Where did we get this number "$\pi/4$"? It has to be an irrational number (in order to ensure that bnd (B) will be the empty set. For example, if B were all (x, y) in X with $x > 1$, then bnd (B) would be al (x, y) in X with $x = 1$ – not the empty set). Furthermore, it has to be less than one (to ensure that some point of A is in B), and also greater than minus one to ensure that some point of A is not in B). So, any irrational between -1 and 1 would do.] Finally, let A consist of all points (x, y) of X with $x = 1$, a "vertical line". This subset of X is connected.

In the metric space of the last figure on page 35, the only connected subsets are those consisting of just one point, and the empty subset.

We turn, finally, to the definition of "bounded" for a set in a metric space. We begin by nothing that our old definition of "bounded", for a set in the plane, can indeed be immediately generalized to a subset of any metric space. It turns out, however, that the resulting definition is not a very convenient one. The reason for this is a rather technical one: One would have to look in detail at various possible applications of this notion of "bounded", and see that it just does not seem to fit properly. Perhaps one example will at least illustrate the idea. Consider the metric space X, d with X the plane, and $d(p, q)$ = geometrical distance between p and q if this is less than one, and one otherwise. We have see that this metric space is "very much like the plane", in that it always gives the same boundaries, the same interiors, and the same connected sets as the plane with usual geometrical distance. Does it give the same bounded sets? Let us try it, using the old definition of bounded (for X, d a metric space, and $A \subset X$, A is bounded if for some point p of X and some positive number c, every point q of A satisfies $d(p, q) \leq c$). Let, for example, $A = X$. Then, if our metric space is the plane with the usual geometrical distance, then A of course will not be bounded. Will A be bounded in this metric space? The answer is yes. Let p be any point of X (say, $(-13, 233)$), and let $c = 2$. The locus of points within distance (using the "funny" d of this metric space) 2 of p is the entire set X. So, every point of A is within 2 of p. So, this A is bounded. Thus, if we carry over directly our old definition of bounded, we will get different answer for what is bounded in the plane with usual geometrical distance and the plane with this distance d.

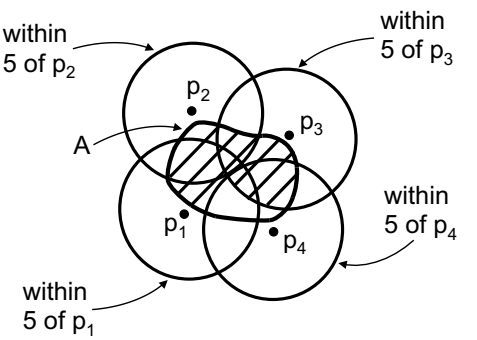
What is the mechanism for this discrepancy? It is, roughly speaking, the following. There is only a difference between our distance d and usual geometrical distance for "larger distances" (namely, those exceeding one). On the other hand, all that really

counts for boundaries, interiors, and connectedness is the "smaller distances". It was because geometrical distance and distance d agree for the "smaller distances" that these two metric spaces yielded the same boundary, interior, etc. On the other hand, what counts for the present definition of "bounded" is, not the "smaller distances", but rather the "larger distances", for, to show boundedness, one wants to choose c very large so that all of A will be within distance c of p. What we would really like, then, is some definition of "bounded" which concerns itself more with "smaller distances". An appropriate definition, as it turns out, is the following.

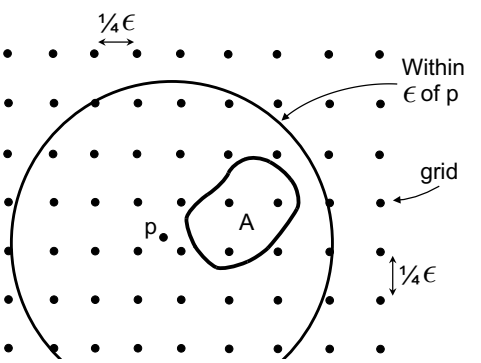
Definition. Let X, d be a metric space, and let $A \subset X$. Then A is said to be *bounded* if, for any positive number ϵ, there is a finite set of points of X, $p_1, p_2 \ldots p_n$, such that, for any point q of A, there is at least one of the p_i's such that $d(p_i, q) \leq \epsilon$.

What the definition requires, in other words, is that, given *any* positive ϵ, one can find a *finite* set of points in X such that every point of A is within ϵ of *at least one* of those points.

It is all very nice to write down a definition – but we had better check that it has at least some correspondence with our intuitive ideas of "bounded'. The most convincing "check" would be to demonstrated that this definition yields exactly the same answers as our earlier definition, at least for the usual plane. The fact is that in this case, it does.



Let X, d be the usual plane. Let A be a subset of the plane, and suppose that A is bounded by the definition above. We shall show that A is bounded by the definition on page 12, i.e., we shall find a point p of X and a positive number c such that every point of A is within c of p. Since A is to be bounded by the definition above, we know that for every positive ϵ, there is a finite set of points of $X \ldots$ In particular, then, this must be true for $\epsilon = 5$. Thus, we have some finite set of points of X, $p_1 \ldots, p_n$, such that every point of A within 5 of at least one of the p_i.. Let us choose $p = p_1$. [Any of the p_i would do.] Then we can compute the distances $d(p, p_2)$, $d(p, p_3)$, and so on to $d(p, p_n)$. These will of course all be just real numbers. Let b denote the *largest* of those real numbers. What we have, then, is that all of $p_1, p_2 \ldots, p_n$ are within distance b of p $(= p_1)$, since b was the largest distance. But every point of A is within distance 5 of one of the p_i. It now follows from the triangle inequality that every point of A is within distance $b + 5$ of p (for given any point q of A, there is some p_i with $d(p_i, q) \leq 5$, while $d(p, p_i) \leq b$, and so, by triangle, $d(p, q) \leq b + 5$). Thus, we can set $c = b + 5$: Every point of A is within distance c of p. In short, we have shown that A is bounded by the definition on page 12.



So, every set in the usual plane bounded according to the definition on the previous page is bounded according to the definition on page 12. We now show the reverse (which is a little bit trickier). Let A be a set in the plane bounded according to the definition on page 12. So, we have a point p and a positive number c such that every point of A is within c of p. We must show boundedness according to the definition on the previous page. Let any positive number ϵ be given. We must find a finite set of points of X such that every point of A is within ϵ of one of them. The method is the following. We introduce a "grid" of points in the plane, consisting of all points (x, y) of the form $n\epsilon/4, = m\epsilon/4)$,

where n and m are integers. This "grid" includes, for example, the points $(13\epsilon/4, -\epsilon/4)$ and $(-7\epsilon/4, 0)$ – it is a rectangular array of points in the plane,spaced $\epsilon/4$ apart. The point of this grid is that, because of the spacing between the points,*every point of the plane* is within distance ϵ of at least one of them. But, unfortunately, this grid will not do the job (for the $p_1, \ldots, p_n$) for the definition on page 45, because there are of course an infinite number of points in the grid. So, we do the following. We consider only the points of our grid which are within distance c of p, i.e., which are within the disk in the figure. This will be a *finite* number of points: call them $p_1, p_2, \ldots, p_n$. [Of course, "n" may be very large, perhaps in the millions, depending on how small ϵ is.] Now clearly every point in our disk, i.e., every point within distance c of p, is within distance ϵ of at least one of the $p_1 \ldots, p_n$ (for every point of the whole plane is within ϵ of *some* point of the original grid, and we obtained the p_i by just throwing away those points of the grid which do not make any difference as far as points in our disk are concerned). But every point of A is within distance c of p. So, we conclude that every point of A is within distance ϵ of at least one of the points $p_1, p_2, \ldots, p_n$. Thus, we began with a set A bounded according to the definition on page 12, chose any positive ϵ, and found a finite set of points, $p_1, p_2, \ldots, p_n$, such that every points of A is within ϵ of at least one of the p_i. We have shown, then, that every set bounded according to the definition on page 12 is bounded according to the definition on page 45.

Putting these together, then, we have shown (although we have not actually written out the formal proof for)

Theorem . Let X, d be the metric space which is the plane with usual geometrical distance, and let $A \subset X$. Then A is bounded according the definition on page 12 if and only if A is bounded according to the definition on page 45.

I hope the discussion above makes it clear what the mechanism of the definition on page 45 is. In order to show that a set is bounded, one must give oneself a positive ϵ, and find a finite number of points of X such that every point of A is within ϵ of at least one of those points. Generally speaking, "the smaller ϵ is, the more points that will be required in the finite set". Note that our new definition of "bounded" deals essentially with "the smaller distances".

The above is an example of a common practice in mathematics. One begins with a definition which seems "reasonable" in a certain context. (Here, "bounded" for sets in the usual plane). One broadens the context (here, to metric space), and then discovers that the old definition does not seem to work quite right in that broader context. So, one looks for a new definition (here, that on page 45) which essentially agrees with the old one in the context in which the old one was applicable (here, the new definition not only "essentially agrees", but "exactly agrees"), but which seems more appropriate in the broader context.

Hereafter, "bounded", for a set in a metric space, always refers to the definition on page 45.

We invented this new definition of "bounded" in part because of a certain example. Let us now go back and see what the new definition does to that example. Let X be the plane, and let $d(p, q)$ be the geometrical distance if it is less than one, and one otherwise. Set $A = X$. Is A bounded (according, of course, to the definition on page 45)? Let us try $\epsilon = 3$. Is there a finite set of points of X such that every point of A (i.e., every point of X) is within 3 of at least one of them? There is: Let p_1, any point be the "finite set of points". The locus of points within 3 of p_1 is all of X, and no every point of A is within 3 of "one of them" (namely, the only one, p_1). But, to show bounded, we have to do this for *every* positive ϵ. What about $\epsilon = 1/2$? Is there a finite set of points in X such that every point of A is within $1/2$ of at least one of them? The locus of points within $1/2$ of a point is a disk of radius $1/2$ centered at that point. Can we find a *finite* number of such disks (all of radius $1/2$) such that every point of X $(= A)$ is in at least one of those disks? Obviously, we cannot. So, we cannot find such a finite set for $\epsilon = 1/2$. So, A in this example is not bounded.

Note how clever the new definition is. It "corrects" the answer in this example, to

yield that the entire set X is not bounded. It does so in a way so that no answers are "corrected" for the usual plane with geometrical distance, and also such that "small distances are the ones which really count". Let X be the plane, and let $d(p, q) = 1$ if $p \neq q$, and zero otherwise. Is $A = X$ bounded? Try $\epsilon = 1/2$. Is there a finite set $p_1 \ldots, p_n$ in X such that every point of X $(= A)$ is within $1/2$ of at least one of the p_i? The locus of points within $1/2$ of a point p consists of just p itself. So, if every point of X is to be within $1/2$ of one of the p_i, every point of X must *be* one of the p_i. But this is impossible, for one is allowed only a finite number of p_i, while X itself has an infinite number of points. So, X itself is not bounded, since we cannot find such a finite set in X for $\epsilon = 1/2$. Clearly, the only subsets of X which are bounded are the finite subsets of X. [Of course, every finite subset of X is bounded. If A, consisting of just points $p_1 \ldots, p_n$, is a finite subset of X, and positive ϵ is given, consider the finite subset of X consisting of A. Then every point of A will be within ϵ of at least one point of this finite set.]

Let X be the set of all disks in the plane, with d given using areas, as before. For $A = X$, A is not bounded. [Try $\epsilon = 1$. Can one find a finite set of disks such that every disk is within distance (using areas) 1 of one of these?] The subset of X consisting of all disks with center the origin is also not bounded. The subset of X consisting of all disks with center the origin is also not bounded. The subset consisting of all discks which lie within a square is bounded (but it is a bit tricky to see).

For the metric space of the last figure on page 35, every subset of X is bounded.

Let X be the points (x, y) of the plane with x rational. For A the subset of X consisting of all (x, y) with $x^2 + y^2 < 1$ and x rational, A is bounded; for A consisting of all (x, y) with $x = 1$, A is not bounded. For $A = X$, A is not bounded.

8. Theorems on Sets in Metric Spaces

We now have the notion of a metric space, and some examples; the definitions of interior, boundary, connected, and bounded in a metric space, and some examples. The next step is the same as the corresponding next step for the plane: We wish to prove some theorems which relate these definitions to each other. One expects, of course, that many of our old theorems for the plane will have analogous versions for metric spaces (with only "typographical changes", such as "point of the plane" becoming "point of X", etc). This feature, after all, was one of our guides in inventing metric spaces. In fact, one often thinks up theorems and their proofs in metric spaces by "thinking plane" (or some other very simple metric space), but "writing metric space". One must be a bit careful in doing this, however, for, when proving a theorem about metric spaces, one is of course only allowed to use those properties of distance which are given in the definition of a metric space. It even turns out on occasion that it is easier to prove something in the more general context of a metric space than in the more special context of the plane. Whereas there are millions of properties of geometrical distance in the plane, there are only three properties of distance in a general metric space. Thus, in the case of metric spaces, there are fewer possibilities for what to do if, in the middle of a proof, one is trying to get from something to something.

Below are various theorems in metric spaces. Most generalize earlier results for the plane. We occasionally omit proofs which are virtually identical to earlier proofs.

Theorem. Let X, d be a metric space, and $A \subset X$. Then int $(A) \subset A$. Proof: Let p be a point of int (A), so, for some positive number ϵ, every point within ϵ of p is in A. But p is within ϵ of p, and so p is in A.

[This is of course the theorem on page 18, but now for metric spaces. The "every point" of the first sentence in the proof means "every point of X", and "within" means "within using distance d of the metric space". That p is within ϵ of p follows from the first condition in the definition of a metric space.]

Theorem. Let X, d be a metric space, and let $A \subset X$. Then int (int (A)) = int (A).

Proof: By the theorem above, every point of int (int (A)) is in int (A). For the converse, let p be a point of int (A). Then for some positive ϵ, every point within ϵ of p is in A. Let q be any point within $1/2\epsilon$ of p. Then every point within $1/2\epsilon$ of q must, by the triangle inequality, be within ϵ of p, and so must be in A. So, q must be in int (A). Since every point within $1/2\epsilon$ of p is in int (A), p is in int (int (A)).

[See theorem on page 20. All we added was "by the triangle inequality".]

Theorem. Let X, d be a metric space, and $A \subset B \subset X$. Then int $(A) \subset$ int (B).

[See the first theorem on page 21.]

Theorem. Let X, d be a metric space, and $A \subset X$ and $B \subset X$. Then int $(A \cap B)$ = int $(A) \cap$ int (B).

[See the second theorem on page 21.]

Theorem. Let X, d be a metric space, and $A \subset X$ and $B \subset X$. Then int $(A) \cup$ int $(B) \subset$ int $(A \cup B)$.

[See theorem on page 23. The proof is that, word or word.]

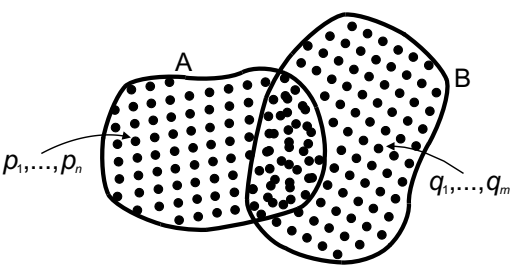
Theorem Let X, d be a metric space, and $A \subset X$. Then bnd A^C) = bnd (A).

[See theorem on page 21,] This, recall, was one of our "interesting" theorems for the plane, for it was suggested by our intuitive ideas of "boundary" and of sets in the plane. There are some pretty strange things which still get to be called metric spaces. Yet, this theorem holds for them all.]

Theorem. Let X, d be a metric space, and $A \subset X$. Then every point of X is in either int A or bnd (A) or int (A^C), and no point of X is in more than one of these sets.

Proof; Let p be a point of X. If for some positive number ϵ every point within ϵ of p is in

A, then p is in int (A). If for some positive number ϵ every point within ϵ of p is not in A, then p is in int (A^C). Clearly, p cannot be in both int (A) and int (A^C). If neither of these holds, i.e., if for every positive number ϵ there is a point within ϵ of p and in A and also a point within ϵ of p and not in A, then p is in bnd (A). Finally, if p is in bnd (A), then (since for every positive ϵ there is a point within ϵ of p and not in A) p is not in int A, and (since for every positive ϵ there is a point within ϵ of p and in A) p is not in int (A^C).
[We did not actually state this as a theorem for the plane, although of course we could have. Note that the second sentence of the theorem could have been "Then int $(A)\cup$ bnd $(A)\cup$ int (A^C) is the set X, and int $(A)\cap$ bnd (A), bnd $(A)\cap$ int (A^C), and int $(A)\cap$ int (A^C) are all the empty set." In the proof, one shows in succession: p can be in int (A); p can be in int (A^C); p cannot be in both; if p is in neither. then it is in bnd (A); if p is in bnd (A), then it can be in neither int (A) nor int (A^C). These assertions, taken together, establish the theorem.]
Theorem. Let X, d be a metric space, and $A \subset B \subset X$. Then, if B is bounded, so is A.
Proof: Let ϵ be a positive number. Then, since B is bounded, there is a finite set, $p_1 \ldots, p_n$ of points of X such that every point of B is within ϵ of one of them. But $A \subset B$, and so every point of A is within ϵ of one of these points. So, A is bounded. [See theorem on page 23. Of course, now we must use the "corrects" definition of bounded, (page 45), and so the present proof is quite different from the earlier proof for the plane. It is reassuring that, even with our new definition of "bounded", we still get this theorem. Note how the present proof begins. We are, after all, trying to show that A is bounded, i.e., that "given any positive $\epsilon, \ldots$". So, we start by giving ourselves a positive ϵ, and we finish by finding a finite set such that every point of A is within ϵ of at least one of the points of that set.]
Theorem. Let X, d be a metric space, and let A and B be bounded subsets of X. Then $A \cup B$ is bounded.
Proof: Let ϵ be a positive number. Since A is bounded, there is a finite set, $p_1 \ldots, p_n$, such that every point of A is within ϵ of one of these points: since B is bounded, there is a finite set, $q_1 \ldots, q_m$, such that every point of B is within ϵ of one of these points. Consider the finite set $p_1 \ldots, p_n, q_1, \ldots, q_m$. Then every point of $A \cup B$, since it must be in either A or B, must be within ϵ of one of these points. So, $A \cup B$ is bounded. [Again, we did not state this one as a theorem for the plane. The present proof is actually a bit simpler than would be the corresponding proof using the old definition of "bounded" in the plane. Of course, one would have expected this result from one's intuitive idea of "bounded". (Score one for the definition.) How does one think of such a proof? One knows that A and B are bounded, so one thinks of a figure such as that on the right, with a "grid" of points such that every point of A is within ϵ of one of them, and similarly for B. One wants to show that $A \cup B$ is bounded, i.e., one must find a new "grid" such that every point of $A \cup B$ is within ϵ of one of these points. What should one choose for the new "grid"? Clearly, one should just combine the grids for A and B. Note that the proof begins "Let ϵ be a positive number."]



Theorem. Let X, d be a metric space, and let A and B be connected subsets of X having some point p in common. Then $A \cup B$ is connected. [The proof is, word for word, that of the theorem on page 23. Again, we have a result which, at least for the plane, reflects our intuitive idea of what "connected" means. But, despite the fact that there are some very strange metric spaces around, it even holds in a general metric space. That theorems such as this one continue to hold gives one confidence that the introduction of a metric space was the right idea.]

Finally, we remark that, since the usual plane is of course a metric space, all our earlier counterexamples for the plane remain counterexamples for metric spaces. Thus,

for example, the following statements (in which "..." stands for "Let X, d be a metric space, and $A \subset X$ and (if B is mentioned) $B \subset X$ are *false*: i) ... No point of A is in bnd (A). ii) ... int (bnd (A)) is the empty set. iii) ... If $A \subset B$, then bnd $(A) \subset$ bnd (B). iv) ... int (bnd $(A \cup B) =$ int $(A) \cup$ int (B). v) ... If A and B are connected, so is $A \cap B$. vi) ... If $A \subset B$ and B is connected, then A is connected. vii) ... If A is connected, then A^c is connected. viii) ... If A is bounded and $A \subset B$ then B is bounded. ix) ... If A is bounded, then A^c is bounded.

9. Topological Spaces

We are now at another crossroads. We have the notion of a metric space, some definitions in metric spaces, some examples, and some theorems which relate the definitions to each other. What should be done next? Shall we invent more definitions, more examples, and more theorems? This would certainly be possible. However, just with the plane before, this is not the direction in which we wish to proceed.

We begin with the observation that there is still something fishy about all this. Recall our discussion of stretching and pulling of the rubber sheet. We decided initially that the kinds of things we were after were those which remained the same under this operation of stretching and pulling. We indeed found four such "things": the interior and boundary, and the notions of connected and bounded. We asked what was genuinely relevant to these notions, and decided that it was distance. This observation was made manifest by expressing interior, boundary, connected, and bounded in a general metric space (i.e., in "a space in which the only thing one has access to is a distance"). What is disquieting, however, is the following fact: Whereas interior, etc. certainly seem to remain the same under the operation of stretching and pulling, "distance" certainly does not. That is to say, the operation of stretching a rubber sheet will certainly change distances between points of the sheet. How does it come to be that, in order to express "stretching-independent" notions, we need something "starching-dependent", namely distance? A more explicit example of this phenomenon is a result from Sect. 7. Let X be the plane, d' the usual geometrical distance, and d the distance with $d(p,q) = d'(p.q)$ if $d'(p,q) < 1$, and $d(p,q) = 1$ otherwise. Then, as we have seen, for any subset A of X one gets the same interior and boundary whether one uses d or d'; the same sets are bounded or connected whether one uses d or d'. These two distances are of course quite different, whereas as far as the structural features we are interested in are concerned, they "might as well be the same". It is as though d and d' different from each other only by a stretching and pulling of the metric space.

The mathematician would regard it as unpleasant that, in order to treat certain features of sets in the plane, one has to invoke structure (namely, a distance) which seems to go beyond what is inherent in the features one is after. It would be taken as the sign that one does not yet understand that one is doing – that one has not yet isolated what is genuinely relevant to "interior", etc. One could hardly doubt that distance has *something* to do with our notions. It is just that it does not seem to be the essential thing – the thing which gives us access to our four notions without forcing one to carry around excess baggage. As the example above makes clear, there is certainly a great deal of "excess baggage" in distance.

So, what is the essential thing? What is *really* relevant here? Here is a problem in mathematics. One might begin, for example, by looking back at what we have done for some clues to the answer. One might in this way stumble across the following clues. First, one might notice that we tend to use our distances in a rather restrictive way. For instance, there never appears, in anything important, a statement such as "The distance is 17.". In fact, the only circumstance in which we actually use our distances is the following. We have some point p. We then say either "for some positive number ϵ" or "for all positive numbers ϵ, and then either "there exists point q with $d(p,q) \leq \epsilon$" or "for all points q with $d)p,q) \leq \epsilon$". It seems as though we do no care about the exact, specific numerical value of $d(p,q)$, but only whether or not it is $\geq \epsilon$, and then, not for a specific numerical value for ϵ, but rather for either "all ϵ "some ϵ". A second clue is an observation we made, and even used, in Sect 7. What seems to count about our distances is, not the larger distances, but only the smaller distances. Thus, for the example on the previous page, we had two distances which "agreed when they are small, but not when they are large".

Theses two distances, however, gave rise to the same interior, etc. But even or the smaller distances, it is not exactly the numerical values which count. For example, let X be the plane, d' usual geometrical distance, and d twice geometrical distance. These distances are of course different – and they are even different for "small distances". However, one easily convinces oneself that d and d' give rise to the same interiors and boundaries, and to the same connected sets and bounded sets.

These clues describe different aspects about what it is we want to retain of our distances. It is "the smaller distances, but even then not exactly their numerical values, but rather their '$\leq \epsilon$ for some ϵ' and '$\leq \epsilon$ for all ϵ' structure". What we want to retain is a sort of shadow of distance which reflects these features.

But the most revealing clue is the following. Consider the fifth theorem on page 48. It says that, given any set A in any metric space, each point of the space is in exactly one of the sets int (A), bnd (A), and int (A^c). But this means the following. Suppose that one were very good at determining interiors, but rather poor at boundaries. Then one could use interiors to determine boundaries. Indeed, because of this theorem, one could determine bnd (A) by finding all points which remain after removing int (A) and int (A^c) from X. Now consider the definition of "disconnected" on page 43. Where does distance enter this definition? At only one place: the "... bnd (B)." at the end (for, of course, distances are needed to determine the boundary). That is to say, if one has determined the boundaries of all sets, then even if one has lost one's sheet giving the distances, one can still determine which sets are disconnected and which are not. We conclude, then, that if one knows the interiors then one can determine the boundaries (without otherwise using the distance), and further that, once one has determined the boundaries, one can determine the disconnected sets (again, without otherwise using the distance). [What of "bounded"? It does not seem to fit into this scheme. We shall return to this shortly.] In short, distance seems to enter this subject essentially only through the determination of interiors.

Here, then, are the clues. The mystery is how to so formulate this subject that the essential things are retained, while the excess baggage is eliminated. Of course, we have very much prejudiced the issue by our choice of clues. In practice, one might find eighty clues, which point in twelve different directions. One would begin with the most promising of these directions, and work one's way along, hoping eventually to find a formulation which seems to have the appropriate character.

The solution to our mystery is the following. What did we do when we passed from the plane to a general metric space? For the plane, we *computed* the distance, using the formula on page 6.

For metric spaces, however, distances were not *computed*: rather, they were just *specified* (i.e., invented), as part of the definition of a metric space. We now want to pass from a metric space to something else. The crucial thing seems to be the interiors. For a metric space, the interiors are *computed* (using the distance in that metric space, and the definition of "interior"). The idea, then, is that in our new spaces, the "interiors will just be specified by fiat, and not computed from anything else". In this way, we shall get rid of those awkward distances, but retain the essential thing we need about the distance, namely the interiors it determines.

But how can one tell, in a metric space, if a given set is the interior of something? That is easy.

Theorem. Let X, d be a metric space, and $A \subset X$. Then there exists a subset B of X such that $A = \text{int}\,(B)$ if and only if $A = \text{int}\,(A)$.

Proof: If $A = \text{int}\,(B)$ for some $B \subset X$, then, by the last theorem on page 48, int $(A) = A$. For the converse, let $A = \text{int}\,(A)$. Then there is a set B (namely, A) such that $A = \text{int}\,(B)$.

Thus, the sets which are the interiors of *something* are just the sets which are the interiors of *themselves*.

Definition. Let X, d be a metric space, and $A \subset X$. Then A will be said to be *self-interior* if $A = \text{int}\,(A)$.

It is the self-interior sets which seem to play the critical role in the study of sets in a metric space.

Recall again what happened in the passage from the plane to the general metric space. We first isolated distance as the essential thing. We next obtained a list of the "distance-like" properties of distance. The final step was to assemble these properties in the definition of a metric space. For the present situation, the essential thing seems to be the self-interior sets. The next step, then, is to figure out what are the "self-interior-like" properties of self-interior sets. With the plane, we were looking for properties of distance which referred only to distance, and not to anything else about the plane. Similarly, we are here looking for properties of self-interior sets which refer only to these sets, and not to anything else about the metric space. Thus, for example, "For A self-interior, there is a set B such that $A = \text{int}\,(B)$." is not appropriate, for it refers to "int", which requires the distance in our metric space. By contrast, "For A self-interior, A^c is self-interior." would be an acceptable property by this criterion, for it refers only to the self-interior sets (and, of course, to the complement, but that is ok, because one does not need the distance to determine the complement). But this property is not a very good one either, for it is false in a general metric space. [For example, let the metric space be the usual plane, and A the usual disk. Then A is self-interior, but A^c is not.] Thus, we are looking for properties which refer only to self-interior, and which are true in a general metric space. Again, there are many more properties than we shall want to retain in the final definition (just as there are many many more properties of distance in the plane than are retained in the definition of a metric space). For example: "For p and q points of X, with $p \neq q$, there is a self-interior set A with p in A and q not in A." is true in any metric space. We shall not retain it. Again, it is a matter of judgment, to select enough properties to capture the essence of what one wants, but not so many that one "practically gets all the way back to a metric space".

The appropriate list of properties, as it turns out, is the following.

Theorem. Let X, d be a metric space. Then

1. X and the empty subset of X are self-interior.
2. For A and B self-interior, $A \cap B$ is self-interior.
3. Given any collection of self-interior sets, their union C is self-interior.

Proof: 1. For p in X, every point within distance 13 of p is in X, and so X is self-interior. The empty set, since it contains no points, is self-interior. 2. Let A and B be self-interior. Then $\text{int}\,(A \cap B) = \text{int}\,(A) \cap \text{int}\,(B) = A \cap B$, where the first equality follows from the first theorem on page 48, and the second equality follows from $A = \text{int}\,(A)$ and $B = \text{int}\,(B)$. 3. By the second theorem on page 48, $\text{int}\,(C) \subset C$. For the converse, let p be a point of C. Then p must be in one of the self-interior sets of the union, say A. Since $A = \text{int}\,(A)$, p is in $\text{int}\,(A)$, i.e., there is a positive number ϵ of p is in A. But $A \subset C$, and so every point within ϵ of p is in C. So, P is in $\text{int}\,(C)$. We conclude that C is self-interior.

Note that, in the third property, we show that the union of *any* collection (possibly two, possibly nineteen, possibly an infinite number) of self-interior sets is self-interior. Of course, the union of an arbitrary collection of sets is the set of all points which are in any one of those sets; the intersection of an arbitrary collection of sets is the set of all points which are in all of those sets. For example, let X be the plane, and consider the collection of subsets of X consisting of all disks (of any positive radius) centered at the origin. That is, for a any positive number, the subset of X consisting of all (x, y) with $x^2 + y^2 < a$ is in our collection. Then the union of this collection is the entire plane (for every point of the plane is in at least one of these disks.) The intersection of this collection is the set consisting of the single point at the origin (for this is the only point which is in every one of our disks). In this example, all the sets in our collection are self-interior in the metric space with d usual geometrical distance. Note that their union, the entire plane, is also self-interior, as demanded by the third property. Their intersection, however, is not self-interior. This example shows, then, that in the second property of the theorem above, we could not have had that the intersection of any collection of self-

interior sets is self-interior. Thus, for self-interior sets, one can take intersections of two, but unions of arbitrary numbers, and still get self-interior sets.

The next step is obvious. Just as the decision as to what are the "appropriate" properties of distance in the plane led to metric spaces, so the decision as to the appropriate properties of self-interior sets leads to a kind of space.

Definition. A *topological space* consists of a set X together with a collection $\mathcal{J}$ of subset of X such that the following conditions are satisfied:

1. X and the empty subset of X are in the collection $\mathcal{J}$.
2. For any sets A and B in the collection $\mathcal{J}$, $A \cap B$ is in the collection $\mathcal{J}$.
3. For any collection of sets in $\mathcal{J}$, their union is also in $\mathcal{J}$.

Thus, to specify a topological space, one must say what the set X is, and what the collection $\mathcal{J}$ of subsets of X is (i.e., which subsets of X are to be in the collection $\mathcal{J}$, and which are not). The subsets in the collection $\mathcal{J}$ are usually called *open* sets (in the topological space X, $\mathcal{J}$), i.e., "open set" is just a short way of saying "subset of X which is in the collection $\mathcal{J}$".

The definition of of a topological space is, in my opinion, one of the most beautiful definitions in mathematics. On the one hand, the definition itself is simple and elegant. [One only has to specify a set X and some subsets, subject to simple conditions. One does not have to give a lot of complicated, detailed information, such as the distances for a metric space.] On the other hand, this single definition gives one access to an enormous and rich range of intuitive ideas – those we have discussed so far as well as many others. The study of topological spaces is a large and important branch of mathematics – called, of course, topology. This will be the subject of our attention hereafter. I hope that we shall be able to give some insight into the power and elegance of this definition.

We give some examples of topological spaces. Just as the plane led immediately to an example of a metric space, so any metric space leads immediately to an example of a topological space. Indeed,

Theorem. Let X, d be a metric space. Denote by $\mathcal{J}$ the collection of all self-interior (using d) subset of X. Then X, $\mathcal{J}$ is a topological space.

Proof: This is immediate from the theorem on 53. Note that, given metric space X, d, we specified the information we needed to specify to have a candidate for a topological space, namely a set (here, X), and a collection of subsets of X (here, all self-interior subsets, using d). We already have the necessary properties, since after all it was just these properties for self-interior subsets of a metric space which suggested our definition of a topological space. The topological spaces so constructed are important enough to deserve a name.

Definition. Let X, d be a metric space. The topological space X, $\mathcal{J}$ obtained by the theorem above will be called the *underlying* topological space of the metric space X, d.

Thus, we instantly obtain many topological spaces, as the underlying topological spaces of our various metric spaces. For example, let X be the plane, and d usual geometrical distance. Then, for the underlying topological space of this metric space, X is of course the set of points of the plane, while $\mathcal{J}$ (the collection of open sets of our topological space) is the collection of self-interior sets of the plane (e.g., the empty set, X itself, any disk, any union of disks, and so on). Next, let X be the plane, but let d be the distance which is geometrical distance if that is less than one, and one otherwise. Then, as we have seen, X, d is a metric space. What is the underlying topological space? The set is of course X, the plane. Now, however, the open sets are to be those self-interior using this d. What, are these sets? Recall that, on page 48, we showed that the interior, using *this* distance d, is always exactly the same as the interior using the usual geometrical distance. Clearly, then, the self-interior sets, in *this* metric space, will be exactly the same as the self-interior sets using the geometrical d. Thus, we obtain exactly the same underlying topological space as we obtained above, for the usual distance in the plane. [How nice! One of the things we were unhappy about was that we seemed, with metric space, to be carrying around excess baggage, reflected by the fact that this distance and geometrical distance

seemed to be "essentially the same as far as the things we are interested in are concerned". The definition of the underlying topological space automatically throws away the excess baggage. We get the same underlying topological space in the two cases.]

Let X be the plane, and let d be given by $d(p,q) = 0$ if $p = q$, and $d(p,q) = 1$ if $p \neq q$. This is a metric space. The set of its underlying topological space is just the plane X. To find the open sets of this topological space, we must find the self-interior sets of our metric space. When, in this metric space, is int $(A) = A$? This is *always* true, for any $A \subset X$, in this metric space. Hence, every subset of X is an open set in this topological space. More generally, we have

Theorem. Let X be a set, and let $\mathcal{J}$ be the collection of all subsets of X. Then X, $\mathcal{J}$ is a topological space.

Proof: 1. X itself and the empty subset are subsets of X, hence in $\mathcal{J}$. 2. The interaction of two subsets of X is a subset of X, hence in $\mathcal{J}$. 3. The union of any collection of subsets of X is a subset of X, hence in $\mathcal{J}$.

Thus, for *any* set X we obtain, via this theorem, a topological space (namely, one in which every subset of X is open). These have a special name.

Definition. A topological space X, $\mathcal{J}$ for which every subset of x is in $\mathcal{J}$ is said to be *discrete*.

Consider the metric space on page 35, in which X consists of just four points, with distances 17 and 1. Its underlying topological space (in which the X, of course, will be just this set with four points) is discrete.

Of course, to give a topological space, one must give nothing less, and need give nothing more, than required by the definition. Every metric space, as we have just seen, gives rise to a topological space – but this need not be the *only* way of obtaining topological spaces. Here is another kind of topological space.

Theorem. Let X be a set, and let $\mathcal{J}$ be the collection of subsets of X consisting of X itself and the empty subset. Then X, $\mathcal{J}$ is a topological space.

Proof: 1. X and the empty set (here written $\emptyset$) are in $\mathcal{J}$. 2. Since $X \cap X = X$, $X \cap \emptyset = \emptyset$, $\emptyset \cap X = \emptyset$, and $\emptyset \cap \emptyset = \emptyset$, the intersection of any two sets in $\mathcal{J}$ is in $\mathcal{J}$. 3. The union of any collection of sets in $\mathcal{J}$ is X if any one of those sets in the union is X, and $\emptyset$ if none of those sets in the union is X. So, any such union is in $\mathcal{J}$.

The topological spaces obtained via this theorem have "as few open sets as possible", namely just X and the empty set. [Clearly, one cannot have have fewer open sets than this, because one has to satisfy the first condition for a topological space.] By contrast, a discrete topological space has "as many open sets as possible", namely every subset of X is open. These two, then, are the extremes for how many open sets there can be. The one with "very few open sets" also has a name.

Definition. A topological space $X, \mathcal{J}$ for which $\mathcal{J}$ consists only of X itself and the empty subset is said to be *indiscrete*.

Just as, in a metric space, "X does not tell one what distance to choose", so, in a topological space, "X does not tell one what open sets to choose". One's only obligation is to specify clearly what subsets of X are or are not in, $\mathcal{J}$, and check the conditions.

Let X be the plane, and let $\mathcal{J}$ consist of all subsets of X having a finite number of points (including the possibility of no points). Thus, the usual disk is not in $\mathcal{J}$, for it has more than a finite number of points. We check the conditions for a topological space. The empty subset of X is certainly in $\mathcal{J}$, for it has no points. However, X itself is not in $\mathcal{J}$, for X does not have a finite number of points. The first condition fails. For the second condition: Is it true that, if each of A and B has a finite number of points, then $A \cap B$ has a finite number of points? This is true, and so the second condition holds. For the third condition: Is it true that, given any collection of subsets of X, each with a finite number of points, their union has a finite number of points? Let us try it for two subsets in the collection. Is it true that, if each of A and B has a finite number of points, then $A \cup B$ has a finite number of points? This is true. However, this third condition fails in general (i.e., when we allow ourselves to take the union of *any* collection of sets in $\mathcal{J}$). Indeed,

consider the collection of all subsets of X having just one point (so the subset consisting of just the origin is in this collection, while a disk is not). This is certainly a collection of sets in $\mathcal{J}$ (for each set in our collection has a finite number of points, namely just one point). What is the union of this collection of sets? It is, we claim, the entire plane X. [Indeed, let p be any point of the plane. Then the subset of X consisting of just p is in our collection. So, p is in the union of the sets in our collection.] But the plane, of course, is not in $\mathcal{J}$. So, we have found a collection of sets in $\mathcal{J}$ whose union is not in $\mathcal{J}$. The third condition fails. Thus, we do not have a topological space (and, in fact, we only satisfied one condition out of the three).

The "mirror image" of the attempt above, however, works. Let X be the plane, and let $\mathcal{J}$ consist of the empty set, X itself, and any subset of X consisting of all of X except for a finite number of points. Thus, for example, the disk is not in $\mathcal{J}$, for it is not the empty set, not X, and it is not all of X except for a finite number of points. But the subset of X consisting of all (x, y) except $(3, -271)$ is in $\mathcal{J}$, for it is all of X except for a finite number of points (namely, except for one point). This $X, \mathcal{J}$, we claim, is a topological space. 1. X and the empty set are certainly in $\mathcal{J}$. 2. Let A and B be in $\mathcal{J}$. If either is the empty set, then $A \cap B$ is the empty set, and so is in $\mathcal{J}$. Suppose, then, that neither is the empty set: Say, A is all of X except the finite number of points $p_1 \ldots, p_n$ of X, and B is all of X except the finite number of points $q_1 \ldots, q_m$ of X. Then $A \cap B$ (the set of all points in both A and B) must be all of X except for points $p_1 \ldots, p_n, q_1 \ldots, q_m$ (for every other point of X is certainly in both A and B, while these points are not in both A and B). But this set $A \cap B$ is all of X except for a finite number of points, and so is in $\mathcal{J}$. 3. Consider any collection of sets in $\mathcal{J}$, and let C be their union. If all the sets in this union are the empty set, then C is the empty set, and so C is in $\mathcal{J}$. Suppose not, so there is some set A in this union which is not the empty set. Then A, since it is in $\mathcal{J}$, must be all of X except for a finite number of points. But C is a union of A with some other sets, and so $A \subset C$. But, since A is all of X except for a finite number of points, and $A \subset C$, C must be all of X except for a finite number of point (for C has to include at least everything in A). So, C is in the collection $\mathcal{J}$. That is, the union of any collection of sets in $\mathcal{J}$ is itself in $\mathcal{J}$. We conclude, therefore, that this is a topological space.

Compare this last example (in which we obtain a topological space) with the preceding one (in which we do not). Why is "finite subsets' of X" so different from "subsets of X consisting of all of X accept for a finite number of points"? One easily traces the reason to the difference between the second and third conditions for a topological space: Whereas the intersection of just *two* open sets must be open, the union of *any collection* of open sets must be open.

Let X consist of three points, p, q, and r. Let $\mathcal{J}$ consist of the empty subset of X, X itself, the subset of X consisting of just p, and the subset of X consisting of just q and r. One checks that this $X, \mathcal{J}$ is a topological space. If, however, $\mathcal{J}$ had consisted of the empty subset of X, X itself, the subset of X consisting of p and q, and the subset of X consisting of q and r, then we would not obtain a topological space: The second condition would fail.

We introduce one final method for obtaining topological space. Recall the following construction for metric spaces. Let X, d be a metric space, and let X' be *any* subset of X. We introduce a distance d' in the set X', as follows. Let p and q be any two points of X'. Then, since $X' \subset X$,, p and q are also points of X. But we already have a distance, d in X. We now *define* our distance in X' to be just this distance in X, i.e., we set $d'(p, q) = d(p, q)$. We indeed have specified a rule which assigns a number (namely, $d'(p, q')$) to any two points, p and q, of X'. We finally note that this X, d' is a metric space. [Each of the three conditions for X', d' to be a metric space follows immediately from the corresponding condition in the metric space X, d.] There is an analogous construction in topological spaces. Let $X\mathcal{J}$ be a topological space, and let $X' \subset X$. Just as above we used the distances in X to obtain distances in X', now we wish to use the open sets in X (i.e., those in the collection $\mathcal{J}$) to obtain our open sets X' (i.e., those which will be in our

collection $\mathcal{J}'$). This we do as follows. Let A be any open set in the topological space X, $\mathcal{J}$. We cannot of course simply say "We shall then regard this A an an open set in X', $\mathcal{J}'$.", for A may not even be a subset of X'. We therefore "force" A to e a subset of X', as follows. We simply take the intersection of A (a subset of X) and X' (also a subset of X). This intersection (of course, a subset of X) will of course also be a subset of X'. It is these subsets of X' (namely, all those obtained by intersecting open sets in X, $\mathcal{J}$ with X') that we wish to deem open in X', $\mathcal{J}'$. With metric spaces, the fact that our X', d was a metric space was immediate from the fact that X, d was a metric space. Similarly, here, one shows that X', $\mathcal{J}'$, constructed as above, is a topological space using the fact that X, $\mathcal{J}$ began as a topological space. That is,

Theorem. Let X, $\mathcal{J}$ be a topological space, and $X' \subset X$. Let $\mathcal{J}'$ denote the collection of all subsets of X' of the form $A \cap X'$, where A is in the collection $\mathcal{J}$. Then X', $\mathcal{J}'$ is a topological space.

Proof: 1. Since the empty set $\emptyset$ is in $\mathcal{J}$, and $\emptyset \cap X' = \emptyset$, the empty set is in $\mathcal{J}'$. Since X is in the collection $\mathcal{J}$, and $X \cap X' = X'$, the set X' is in $\mathcal{J}'$. 2. Let A' and B' be in $\mathcal{J}'$, say $A' = A \cap X'$ and $B' = B \cap X'$, with A and B in the collection $\mathcal{J}$. Then $A' \cap B' = (A \cap B) \cap X'$, while, since X, $\mathcal{J}$ is a topological space, $A \cap B$ is also in $\mathcal{J}$. Hence, $A' \cap B'$ is in $\mathcal{J}'$. 3. Let C' be the union of any collection of sets in $\mathcal{J}'$. Then each of the sets in this union is the intersection, with X', of a set in the collection $\mathcal{J}$. Denote by C the union of the corresponding sets in $\mathcal{J}$. Then, since X, $\mathcal{J}$ is a topological space, this C is also in the collection $\mathcal{J}$. But we have $C' = C \cap X$. Hence, C' is in $\mathcal{J}'$.

In this proof, we have used a few (easy) facts about intersections and unions. [For example, in the second step, we used the fact that if one takes the intersection of $A \cap X'$ with $B \cap X'$, the result is the same as taking the intersection of $A \cap B$ with X'. In the third step, we used the fact that the result of taking the union of a collection of sets of the form $A \cap X'$ is the same as the result of first taking the union of all the A's, and then intersecting the result with X'. One checks these facts by the usual, straightforward, arguments. (E.g., "Let p be a point of .... Then p must be in ... or ... but not .... Therefore, p must be in ... etc.")] Using this theorem, then, we again obtain an enormous variety of topological spaces. One begins with *any* topological space, X, $\mathcal{J}$. One takes *any* subset X' of X. Then, "restricting" the open sets of X, $\mathcal{J}$ to X' (i.e., taking their intersections with X'), we obtain a new topological space, X', $\mathcal{J}'$. These topological spaces have a special name.

Definition. Let X, $\mathcal{J}$ be a topological space, and $X' \subset X$. Then the topological space X', $\mathcal{J}'$, obtained via the theorem above, is called the *topological subspace* (of X, $\mathcal{J}$, based on the subset X').

To summarize, any set X gives rise to two topological spaces (the indiscrete one and the discrete one); any metric space gives rise to a topological space (the underlying one); any metric space gives rise to a topological space (the underlying one); any subset of any topological space rise to a topological space (the topological subspace).

10. Interior, etc. in Topological Spaces

We initially introduced the notion of the interior, boundary, connected, and bounded for sets in the plane. These notions satisfied certain properties. We then introduced a sort of "generalized plane", a metric space. We were able to recover, in metric spaces, our four notions, with similar properties. We have now introduced "generalized metric spaces", i.e., topological spaces. We now wish to recover our four notions again in a general topological space, and again find their properties.

In our discussion of sets in the plane, we constantly made use of "disks of various radii centered at various points". In metric spaces, the analogous things were the subsets of X given by "the locus of all points q of X with $d(p,q) < \epsilon$". Subsets of this form were used in virtually every argument for metric spaces. In particular, these subsets of a metric space reduce, for the case of the plane, to our usual disks. There is a similar class of subsets for a topological space, which subsets will play very much the same role. These are the following.

Definition. Let X, $\mathcal{J}$ be a topological space, and p a point of X. An *open neighborhood* of p is an open set in this topological space, containing the point p.

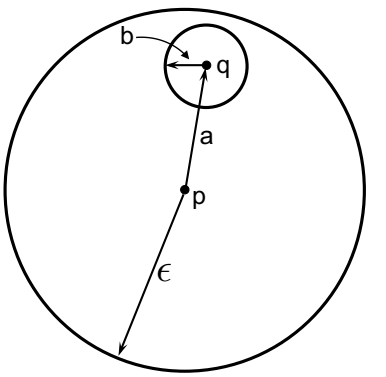
We intend to use open neighborhood in topological space very much as we used disks in the plane, and "the locus of points q with $d(p,q) < \epsilon$" in a metric space. Is this at all feasible? Do open neighborhoods bare any resemblance to "the locus of points q with $d(p,q) < \epsilon$"? We have to ask this question with a little bit of care, We cannot, for example, just go off and ask "Are the open neighborhoods just given by the locus of points $q \ldots$?". The problem here is that open neighborhoods are defined in *topological spaces*, while "the locus of points $q \ldots$" are defined in *metric spaces*. But topological spaces and metric spaces are quite different things. We do not have any distance in a topological space, and we do not have any open sets in a metric space. But there is, fortunately, one area of common ground, on which we can check things out. We know that every metric space gives rise to a topological space, its underlying topological space. Thus, we may begin with a metric space, X, d, then determine its underlying topological space $X, \mathcal{J}$, and finally ask for the comparison between "the locus $\ldots$" in the *metric space* X, d and the open neighborhoods in the *topological space* $X, \mathcal{J}$. [Of course, this will just represent a rough check that we are on the right track. There are, as we have seen, topological spaces which do not arise from *any* metric space. Nonetheless we permit ourselves the definition "open neighborhood" in *every* topological space.] The comparison results in the following.

Theorem. Let X, d be a metric space, $X, \mathcal{J}$ its underlying topological space, p a point of X and ϵ a positive number. Let A be the set of all points q of X, with $d(p,q) < \epsilon$. Then, in the topological space $X, \mathcal{J}$ A is an open neighborhood of p.

Proof: Since $(p,p) = 0$, which is certainly less than ϵ, p is in A. So, we need only show that A is open in the topological space $X, \mathcal{J}$, i.e., that A is self-interior in the metric space X, d. Since, by an earlier theorem, int $(A) \subset A$, we have only to show that every point of A is in int (A). Let q be any point of A, and set $d(p,q) = a$, so $a < \epsilon$. Choose, by this last inequality, positive number b such that $a + b < \epsilon$. Now let r be any point of X with $d(q,r) \leq b$. Then, by the triangle inequality (since $d(p,q) = a$ and $d(q,r) \leq b$), we have $d(p,r) \leq a + b$. But $a + b < \epsilon$, and so $d(p,e) < \epsilon$, and so r is in A. So, we have shown that every point within distance b of point q is in A; hence, that q is in int (A). Since every point of A is in int (A), we have $A =$ int (A).

This is a slightly tricky proof. We are trying to show that A is an open neighborhood of p, and A is defined using the metric space. What we must show, then, is that p in A

(which is easy), and that A is open (which is more difficult). But "open" in our topological space means "self-interior" in the metric space. It all reduces, therefore, to showing that every point of A is in int (A). For this, we select any point q of A, and proceed to show that this q must be in int (A), i.e., proceed to find a positive to show number b such that every point within b of q is in A.



Note, however, that a "b" which works depends on what q is. One could have guessed that this would happen. Recall, way back, our demonstration that, for the disk in the plane, the interior of this disk is just the disk. There, we picked any point of the original disk, and found a smaller disk, centered at that point, which lied entirely within the original disk. Think of p as the center of the original disk, and of q as some point in that disk. How small must our "smaller disk" be, to ensure that it will be a subset of the original disk? Clearly, "how small" depends on what q is. The radius, b, of the smaller disk must be such that $a + b < \epsilon$. This observation about the plane, then, suggests the choice of b in the proof above.

The theorem above shows that "the locus ..." in a metric space always yields an open neighborhood in the underlying topological space. But what about the reverse? How do we pass from open neighborhoods in the underlying topological space to "the locus ..."? The other half of the comparison is the following.

Theorem. Let X, d be a metric space, $X, \mathcal{J}$ its underlying topological space, and p a point of X. Let A be an open neighborhood of p in the topological space $X, \mathcal{J}$. Then there exists a positive number ϵ such that every point q of X with $d(p, q) < \epsilon$ is in A.

Proof: Since A is an open neighborhood of p, p is a point of A. Further, A is open, i.e., A is self-interior in the metric space X, d, i.e., $A =$ int (A). So, since p is in A, p is in int (A). That is, there is a positive number ϵ such that every point q with $d(p, q) \leq \epsilon$ is in A. So, every q with $d(p, q) \leq \epsilon$ is certainly in A. This half is somewhat easier.

The situation, to summarize, is the following. We begin with metric space X, d and find its underlying topological space, $X, \mathcal{J}$. Fix point p of X. Then, for *any* positive ϵ, "the locus of points q of X with $d(p, q) < \epsilon$" (something defined using the metric space) is always an open neighborhood of p in the topological space. The reverse is that, given any open neighborhood of p in the topological space, there is *some* positive ϵ such that every point within ϵ of p is in that open neighborhood (but, of course, the neighborhood need not *be all* points q with $d(p, q) < \epsilon$ – it just has to *include* all such points). Using these two theorems, then, we can pass back and forth between "the locus ..." and "open neighborhood" – provided we are dealing with a metric space and its underlying topological space. [Of course, if we just dealing with a topological space – with no metric space around – then we do not want to "pass back and forth", since we do not have any metric space to "pass" to: We cannot even say "the locus ...".]

We turn now to our first definition – that of "interior" in a topological space. Recall, (page 38) the definition of "interior" for a metric space. What it says, essentially, is that point p gets to be in int (A) if, for some positive number ϵ, "the locus of points q with $d(p, q) < \epsilon$" is a subset of A. But our "generalized locuses", for a topological space, are the open neighborhoods. This immediately suggests how we should make our new definition.

Definition. Let $X, \mathcal{J}$ be a topological space, and $A \subset X$. Then the *interior* of A is the set of all points p of X such that some open neighborhood of p is a subset of A.

The first thing to note about this definition is that it makes sense. It does not refer to distances, i.e., to things we have only in metric spaces. Rather, it refers to the one thing we have access to in a topological space, namely, the open sets (or, what is practically the same thing, the open neighborhoods).

Recall what happened in the passage from the plane to metric spaces. We first defined the "interior" for a set in the plane. Then, once we got metric spaces, we defined "interior" for any subset of any metric space. One example of a metric space is of course the plane, with usual geometrical distance. Thus, we could ask ourselves the following question. Let A be a set in the plane. Then, using the definition on page 11, we can determine the interior of this set in the plane. But we could also do things in a somewhat different way. We might, alternatively, regard our plane, with geometrical distance, as a metric space. Then, we have a metric space X, d (with X the plane, and d geometrical distance), and we have $A \subset X$. Now, we can turn to the definition on page 38, and determine the interior of this subset of a metric space. The question is: Do we obtain the same "interiors", doing it in the two ways? The answer, of course, is that we do, as seen at the top of page 38. Now, we are passing from metric spaces to topological spaces. We can ask a similar question. Suppose we have a metric space X, d, and $A \subset X$. Then we can compute the interior of A, using the definition on page 38. However, we can alternatively do things in a different way. We can take the underlying topological space of our metric space. Then, we have X, $\mathcal{J}$, a topological space, and $A \subset X$. Now, using the definition on the previous page, we can determine the interior of A. Question: Do we obtain the same answer as we got from the definition on page 38. That is, do we obtain the same "interior" if we regard A as a subset of the metric space, and use the definition for a metric space, as if we regard A as a subset of the topological space, and use the definition for a topological space? If not, things would get pretty confusing, because we would always have to specify "which interior". [In fact, in practice, if we could obtain different answers by the two methods, one would probably use a different word for the definition on the previous page, say "interior".] Fortunately, things work out.

Theorem. Let X, d be a metric space, and X, $\mathcal{J}$ its underlying topological space, and $A \subset X$. Then the interior of A in the metric space X, d is the same set as the interior of A in the topological space X, $\mathcal{J}$.

Proof: Write $\text{int}_d\ (A)$ and $\text{int}_{\mathcal{J}}\ (A)$ for the two interiors. Let p be any point of $\text{int}_d\ (A)$, so, for some positive number ϵ, every point q with $d(p, q) \leq \epsilon$ is in A. That is, denoting by B the set of all q with $d(p, q) < \epsilon$, B is a subset of A. But, by the theorem on page 58, B is an open neighborhood of p. So, p is in $\text{int}_{\mathcal{J}}\ (A)$. For the converse, let p be any point of $\text{int}_{\mathcal{J}}\ (A)$, so there is an open neighborhood C of p, with $C \subset A$. By the theorem on page 59, there is a positive number ϵ such that every q with $d(p, q) < \epsilon$ is in C. So, every q with $d(p, q) \leq \epsilon/2$ is in C. But $C \subset A$, and so every q with $d(p, q) \leq \epsilon/2$ is in A. That is q is in $\text{int}_d\ (A)$.

Note how, in the proof, one just uses the two theorems which relate "open neighborhood" with "the locus of points ...". The "$\epsilon/2$ – business" near the end of the proof is necessary to switch from "$<$" (which comes from the theorem on page 59) to "$\leq$" (which is needed in the definition on page 38). One does occasionally have to make such switches, and it is always done in this way (divide by 2).

What this theorem says, then, is that our definition of "interior" on page 59 is a reasonable one. The thing there defined deserves to be called "interior", for it is a genuine generalization of "interior" for metric spaces in the sense that, when both definitions are applicable (i.e., when we work with the underlying topological space of a metric space), they agree. For X, $\mathcal{J}$ *any* topological space, and $A \subset X$, we shall write $\text{int}\ (A)$ for the interior of A. [The use of the same word, "interior", and the same symbol, "int ( )", for metric and topological spaces leads to no confusion, just as there was none in the passage from the plane to a general metric space.]

The theorem on the previous page produces instantly a long list of examples of interiors for sets in various topological spaces. Consider any of our earlier examples of metric spaces. Take the underlying topological space. Let $A \subset X$. Then $\text{int}\ (A)$ (regarding A as a subset of this space) is just $\text{int}\ (A)$ (regarding A as a subset of the metric space). For instance, let X be the plane, and $d(p, q) = 0$ if $p = q$ and 1 otherwise. Then, by the theorem on page 39, we have $\text{int}\ (A) = A$ for any subset A of this metric space. Now consider

the underlying topological space. By the theorem on the previous page, we must have int $(A) = A$ (now, using the definition for "interior" on page 59) for any subset A of this topological space.

We give a few more examples.

Let X be any set, and $X, \mathcal{J}$ the discrete topological space. [That is *every* subset of X is deemed an open set in this topological space.] What are the open neighborhoods in this topological space? An open neighborhood of a point is an open set containing the point. Since every subset of X is open, an open neighborhood of p is any subset of X containing the point p. In particular, the subset of X consisting of p alone is an open neighborhood of p in this topological space. Now let $A \subset X$. We wish to determine the interior of this subset of this topological space. Suppose first that p is a point of A. Then, we claim, p is in int (A), that is, there is an open neighborhood of p which is a subset of A. Indeed, consider the open neighborhood of p consisting of p itself. This one is certainly a subset of A (since p is in A). So all the points of A are in int (A). What about points not in A? These can never be in int (A), and in fact this is true for *any* subset of *any* topological space.

Theorem. Let $X, \mathcal{J}$ be a topological space, and $A \subset X$. Then int $(A) \subset A$.

Proof: Let p be a point of int (A), so there exists an open neighborhood C of p with $C \subset A$. But p is in C and $C \subset A$. But p is in C and $C \subset A$, and so p is in A.

[Compare with the first theorem on page 48 Obviously, we are using virtually the same proof, essentially replacing "locus of points $q \ldots$" by "open neighborhood". It should be emphasized, however, that the theorem above is not an immediate consequence of that on page 48 One is for metric spaces, using the definition on page 38; the other for topological spaces, using the definition on page 59. Not all topological spaces come from metric spaces.] In any case, we conclude that, for A any subset of the topological space under consideration (that introduced on the previous page) int $(A) = A$. This, of course, is the answer we would have expected, from the theorem on page 60. Recall that for the metric space with distance "zero if the points are equal, one otherwise", the underlying topological space is discrete. But in such a metric space, int $(A) = A$ always; hence, also for the underlying topological space.

Let X be any set, and $X, \mathcal{J}$ the indiscrete topological space. [That is, the only subsets of X deemed open are the empty subset and X itself.] What are the open neighborhoods in this topological space? An open neighborhood of p is an open set containing p. But the only open sets are the empty set and X itself, while the empty set does *not* contain the point p. So, in this topological space, there is just one open neighborhood of a given point p, namely X itself. Now let $A \subset X$. We determine int (A). Given a point p of X, when will there be an open neighborhood of p which is a subset of A? But this is easy, since the only open neighborhood of p is X itself. If $A = X$, then clearly our open neighborhood of P will be a subset of A – for *every* point p of X. Thus, we have int $(X) = X$. Suppose, however, that A is not all of X. Then there *cannot* be an open neighborhood of p which is a subset of A (since the only open neighborhood of p is X itself, and this is not a subset of A). So, no point p will get to be in int (A). So, int (A) will be the empty set. Thus, for this topological space, int (A) is X if $A = X$, and the empty set otherwise. [Note that int (A) is always a subset of A, as required by the theorem on the previous page.]

We next establish some theorems for interiors. Many of these will be just the "topological versions" of earlier results for metric spaces. We emphasize again, however, that each theorem must now be proved anew, using the definition of a topological space, and of the interior of a subset of a topological space. But, of course, one uses the proofs for metric spaces as guides.

Theorem. Let $X, \mathcal{J}$ be a topological space, and $A \subset X$. Then int (A) is given by the union of all open subsets of A.

Proof: Point p is in int (A) if and only if p is a point of an open subset of A.

[I hope the proof is clear. Point p being in int (A) just means that it is a point of an open subset of A. But, clearly, the set of points which are points of open subsets of A is just

the set of points which are in the union of open subsets of A.] Note that "the union of all open subsets of A" is, by the third condition for a topological space, an open set, and this is of course a subset of A. Thus, what this theorem means is that int (A) is the largest open subset of A. Thus, to determine int (A), all one has to do is look around for open subsets of A, and find the largest one.

Theorem. Let X, $\mathcal{J}$ be a topological space, and $A \subset X$. Then A is open if and only if int $(A) = A$.

Proof: Let int $(A) = A$. Then, by the theorem above, int (A) is open, and so A is open. Let A be open. Then any point p of A has an open neighborhood which is a subset of A (namely, A itself), and so $A \subset$ int (A). But by the theorem on page 61, int $(A) \subset A$. So, int $(A) = 1$

[This is a bit strange. We motivated our definition of a topological space by looking for properties of self-interior sets in a metric space, and then demanding those properties for a collection of subsets of X in the definition of a topological space. Again using the analogy with metric spaces, we defined the interior for a subset of a topological space. So, we end up with topological spaces, and interiors therein. Nothing stops us, then, from looking at the "self-interior" sets in a topological space. When we do so, we find out that they turn out to be exactly the open sets in our topological space! Note that this theorem does *not* follow directly from any of the things about metric spaces. The definitions for topological spaces stand on their own, and things must be proven therein afresh.]

Theorem. Let X, $\mathcal{J}$ be a topological space. Then the interior of X is X, and the interior of the empty set is the empty set.

Proof: Since X and the empty set are open, by the first condition in the definition of a topological space, this is immediate from the preceding theorem.

Theorem. Let X, $\mathcal{J}$ be a topological space, and $A \subset X$. Then int (int (A)) = int (A).

Proof: By the theorem on page 61, int (A) is open. So, by the first theorem on page 62, int (int (A)) = int (A).

[Compare, the second theorem on page 48. For the corresponding theorem in metric spaces, we used the triangle inequality. It is as though the triangle inequality is somehow "hidden" in our conditions for a topological space. It seems remarkable that this should be possible, for the triangle inequality seems so "numerical", while the conditions for a topological space are so "set-theoretic". But the definition of a topological space clearly retains enough of the flavor of the triangle inequality to recover such results.]

Theorem. Let X, $\mathcal{J}$ be a topological space, and $A \subset B \subset X$. Then int $(A) \subset$ int (B).

Proof: Let p be a point of int (A), so there is an open neighborhood C of p with $C \subset A$. But, since $A \subset B$, this open neighborhood C of p also satisfies $C \subset B$. So, p is in int (B). [Compare, first theorem on page 48.]

Theorem. Let X, $\mathcal{J}$ be a topological space, and $A \subset X$ and $B \subset X$. Then int $(A \cap B)$ = int $(A) \cap$ int (B).

Proof: Let p be a point of int$(A \cap B)$, so there is an open neighborhood C of p with $C \subset A \cap B$. But, since $C \subset A \cap B$, we have $C \subset A$ (and so p is in int (A)), and $C \subset B$ (and p is in int (B)). Hence, p is in int$(A) \cap$ int (B). For the converse, let p be a point of int $(A) \cap$ int (B). Then p is in int (A), and so there is an open neighborhood C_1 of p with $C_1 \subset A$; and p is in int (B), and so there is an open neighborhood C_2 of p with $C_2 \subset B$. Set $C = C_1 \cap C_2$, so certainly point p is in C, and $C \subset A \cap B$. But, by the second condition for a topological space, C is open. Thus, C is an open neighborhood of p, which $C \subset A \cap B$. So, p is in int $(A \cap B)$.

[Compare, the second theorem on page 48, whose proof is on page 21. Note that the proof of the analogous result for metric spaces requires that one take the smaller of two positive numbers, ϵ_1 and ϵ_2. In the topological case, these ϵ's get replaced by open neighborhoods, C_1 and C_2. Instead of taking the smaller of the ϵ's, we take the intersection of C's. The result will again be an open neighborhood, by the second condition for a topological space. Again, we have managed to replace a "numerical operation" (taking the smaller of two numbers) by a "set-theoretic operation" (taking the intersection of two sets). The

definition of a topological space retains enough of the flavor of "comparison of distances" to permit us to recover this result.]

Theorem. Let X, $\mathcal{J}$ be a topological space, and $A \subset X$ and $B \subset X$. Then int $(A) \cup$ int $(B) \subset$ int $(A \cap B)$.

Proof: Since int $(A) \subset A$ and int $(B) \subset B$, int $(A) \cup$ int $(B) \subset A \cup B$. That is, int$(A) \cup$ int (B) is an subset of $A \cup B$. The result now follows from the theorem on page 61.

[Compare, theorem on page 48.]

We turn nest to the definition of "boundary" for a subset of a topological space. As usual, we use metric spaces as a guide. The definition of "boundary" for a subset of a metric space is on page 38. essentially, all we need do is replace "locus of points q with $d(p,q)\epsilon$" by "open neighborhood of p".

Definition Let X, $\mathcal{J}$ be a topological space, and $A \subset X$. Then the *boundary* of A is the set of all points p of X such that, for every open neighborhood C of p, there is a point of C in A, and also a point of C not in A.

Note that the definition even become a bit easier to state for topological spaces.

Immediately after defining "interior" for topological spaces, we carried out a check on the definition. We showed that, *for the case in which the topological space is the underlying one of a metric space*, the definition reduces to that of "interior" in a metric space. We can carry out a similar check for "boundary".

Theorem. Let X, d be a metric space, X, $\mathcal{J}$ its underlying topological space, and $A \subset X$. Then the interior of A in the metric space X, d is the same set as the boundary of A in the topological space X, $\mathcal{J}$.

Proof: Write bnd$_d$ (A) and bnd$_{\mathcal{J}}$ (A) for the two boundaries. Let p be a point of bnd$_d$ (A), so, for every positive number ϵ, there is a point q of A with $d(p,q) \leq \epsilon$, and also a point q' not in A with $d(p,q') \leq \epsilon$. Let C be any open neighborhood of p. Then, by the theorem on page 59, there is a positive number ϵ such that every point r with $d(p,r)\epsilon$ is in C. But, since p is in bnd$_d$ (A), there is a point q in A with $d(p,q) \leq \epsilon/2$ (hence, with $d(p,q) < \epsilon$, hence, with q in C), and also a point q' not in A with $d(p,q') \leq \epsilon/2$ (hence, with $d(p,q') < \epsilon$, hence, with q' in C). So, for every open neighborhood C of p, there is a point of C in A, and also a point of C not in A. So, p is in bnd$_{\mathcal{J}}$ (A). For the converse, let p be a point of bnd$_{\mathcal{J}}$ (A). Let ϵ be any positive number, and let C be the set of points q with $d)p,q) < \epsilon$. Then, by the theorem on page 58, C is an open neighborhood of p. Since p is in bnd$_{\mathcal{J}}$ (A), there is a point q of C (i.e., with $d(p,q) < \epsilon$, hence, with $d(p,q) \leq \epsilon$) in A, and also a point q' of C (i.e., with $d(p,q') < \epsilon$, hence with $d(p.q') \leq \epsilon$) not in A. So, for every positive number ϵ, there is a point within ϵ of p and in A, and also a point within ϵ of p and not in A. So, p is in bnd$_d$ (A).

It is just like the similar result for "interior". We again use our two theorems, on page 58 and 59, to convert from "the locus of point $q \ldots$" to "open neighborhood".

Again, we instantly obtain a long list of examples of boundaries of sets in topological spaces. We begin with any of our metric spaces, choose a subset A, and take the underlying topological space. Then the interior of A in this topological space is exactly the same set at the boundary of A in the original metric space. Again, we denote by bnd (A) the boundary of a subset of a topological space.

We give a few more examples.

Let X be a set, and X, $\mathcal{J}$ the discrete topological space (every subset of X open). Let $A \subset X$. In order that point p be in bnd (A), it must be the case that every open neighborhood of p contain a point of A, and also a point not in A. But one open neighborhood of p is the set consisting of p itself (certainly open in this topological space, and certainly containing the point p). But *this* open neighborhood C could hardly contain a point of A *and* a point not in A, for C only has one point in it. So, no point of X satisfies the condition for being in bnd (A). So, in this case, bnd (A) is the empty set. Again, compare with the theorem on page 39.

Next, let X be any set, and X, $\mathcal{J}$ the indiscrete topological space. Let $A \subset X$. In order that p be in bnd (A), it must be the case that every open neighborhood of p contain a point

of A, and also a point not in A. But, in this topological space, the only open neighborhood of p is the entire set X. So, in order that p be in bnd (A), there must be a point of X in A, and also a point of X not in A. Suppose first that A is the empty set. Then there will never be a point of X in A. So, every point p will fail to satisfy the condition for being in bnd (A). So, bnd (A) will be the empty set. Next, suppose that A is all of X. Then there will never be a point of X not in A. So, every point p will fail to satisfy the condition for being in bnd (A). So, bnd (A) will be the empty set. Finally, suppose that A is neither the empty set nor X. Then there will always be a point of X in A (since A is not empty), and always a point of X not in A (since A is not all of X). So, *every* point p of X will satisfy the conditions for being in bnd (A). So, bnd (A) in this case will be X. Hence, for the indiscrete topological space, bnd (X) and the boundary of the empty set are empty, while the boundary of any other subset is X.

We give one final example of the definitions of "interior" and "boundary". Let $X, \mathcal{J}$ be the topological space introduced on pages 56 and 56 (in which X is the plane, and open sets are X, the empty set, and subsets of X consisting of all of X except for a finite number of points). Let p be a point of X. What are the open neighborhoods of p? The empty set, since it does not contain p, will never do. The set X is of course an open neighborhood of p. Furthermore, all of X except for a finite number of points (provided, of course, that the point p was not left out by being in that finite number!) will be an open neighborhood of p. So, we have the open neighborhoods. Now, let $A \subset X$. Let us try to find int (A). Point p will be in int (A) provided there is an open neighborhood of p. (i.e., all of X, or all of X except for a finite number of points) which is a subset of A. Suppose, for example, that A is our usual disk. Then no point p will satisfy this criterion (for A is just a little old disk: X itself is certainly is not a subset of A, and X except for a finite number of points is not a subset of A either). So, for this A, int (A) is empty. Clearly, int (A) will always be the empty set unless int (A) is "large enough" to contain one of these open neighborhoods. That is, int (A) will be empty unless A is "at least as large as all of X except for a finite number of points". Suppose, then, that set A is all of X except for a finite number of points. Then, of course, A is open in this topological space. But in this case, by the first theorem on page 62, int $(A) = A$. So, in this topological space, int (A) is the empty set unless A consists of all of X except for a finite number of points (for "smaller" A's than these do not contain any open neighborhoods as subsets), while, if A does consist of all of X except for a finite number of points, int $(A) = A$ (for A is then open). Next, we determine some boundaries. In order that p be in bnd (A), it must be the case that every open neighborhood of p contains a point in A and also a point not in A. The only open neighborhoods of p are X itself and all of X except for a finite number of points (where p is not one of that "finite number" left out). Let, for example, A be our disk. Then obviously every one of these open neighborhoods will contain a point in A and also a point not in A (for the open neighborhood must be "practically all of X"). So, for A the disk, bnd (A) will be X (for every point of X will satisfy the condition). Suppose, however, that A is somewhat "larger", say all of X except for a finite number of points, $p_1 \ldots, p_n$. Let p be a point not one of the p_i. Will it be true that every open neighborhood of p contains a point in A and also a point not in A? It will not be true: A itself is an open neighborhood of p, while A certainly does not contain a point not in A. So, the points of A will not be in bnd (A). However, let p be one of the p_i's, say $p = p_1$. Will it be true that every open neighborhood of p contains a point in A and also a point not in A? It will be true: Given any open neighborhood of p, it will certainly contain a point in A (since A is all of X except for a finite number of points, while any such open neighborhood must also be all of X except for a finite number of points). But such a neighborhood will also contain a point not in A, namely the point p itself. So, if p is one of the p_i, then p is in bnd (A). In this example, then, bnd (A) excludes all points in A, but includes the points $p_1 \ldots, p_n$. In other words, bnd $(A) = A^C$. Suppose, next, that A consists of just a finite collection of points, say $p_1 \ldots, p_n$. Let point p be none of the p_i. Is it true that, for every open neighborhood of p, it contains a point of A and also a point not in A? It is not true.

Let C be all of X except for $p_1 \ldots, p_n$. This is an open neighborhood of p. But it contains no point in A (for we left out of C just the n points of A). So, this point: p is not in bnd (A). Let p be one of the p_i, say $p = p_1$. Now, every open neighborhood of p certainly contains a point in A (since the open neighborhoods are practically all of X, while "not in A" is also practically all of X), while every such open neighborhood also contains a point in A (namely p itself). So, this point p is in bnd (A). Thus, in this case, bnd $(A) = A$. To summarize, for this topological space, bnd (A) is A^C if A is all of X except for a finite number of points, bnd (A) is A if A is a finite subset of X, and bnd (A) is X otherwise.

We now obtain a few theorems involving boundaries.

Theorem. Let $X, \mathcal{J}$ be a topological space, and $A \subset X$. Then bnd (A) = bnd (A^C).

Proof: Point p is in bnd (A) if and only if every open neighborhood of p contains a point in A and also a point not in A. But points in A are just those not in A^C. So, point p is in bnd (A) if and only if every open neighborhood of p contains a point not in A^C and also a point in A^C, i.e., if and only if p is in bnd (A^C).

[Compare, the theorem on page 48, proven on page 21. It is almost exactly the same.]

Theorem. Let $X, \mathcal{J}$ be a topological space, and $A \subset X$. Then every point of X is in int (A) or bnd (A) or int (A^C), and no point of X is in more than one of these sets.

Proof; Let p be a point of X. If some open neighborhood of p is a subset of A, then p is in int (A). If some open neighborhood of p contains no points of A, then p is in int (A^C). Clearly, p cannot be in both int (A) and int (A^C). If neither of these holds, i.e., if every open neighborhood of p contains a point not in A and also a point in A, then p is in bnd (A). Finally, if p is in bnd (A), then (since every open neighborhood of p contains a point not in A) p cannot be in int (A), and (since every open neighborhood of p contains a point not in A^C) p cannot be in int (A^C).

[Compare, the theorem on page 48.]

These two results, at least for the plane, seemed "geometrically reasonable", in that they fit in with our intuitive ideas of the words "boundary" and "interior". But they continue to hold even in topological spaces! Again, the definition of a topological space has somehow managed to capture our intuitive ideas.

As an example of these two theorems, consider the topological space discussed on page 64 ("all of X except for a finite number of points"). Let $A \subset X$ be a finite set. Then bnd $(A) = A$, as we saw. Now, A^C is all of X except for this finite set of points. So, as we saw above, bnd A^C) is that finite set, i.e., bnd $(A^C) = A$. That is, we have bnd (A) = bnd (A^C), as demanded by the theorem above. In this example, int (A) is the empty set, and int (A^C) is A^C. So, int (A) = empty, bnd $(A) = A$, and int $(A^C) = A^C$. It is indeed true, as guaranteed by our theorem on the previous page, that every point is in one of these three sets, and no point is in more than one. Alternatively, for A the disk, int(A) = empty, bnd $(A) = X$, and int (A^C) = empty. Again, every point of X is in one and only one of these sets.

Finally, we turn to "connected" for topological spaces. We can just take over directly the definition for metric spaces, page 43

Definition. Let $X, \mathcal{J}$ be a topological space, and $A \subset X$. Then A is said to be *disconnected* if there exists a subset B of X such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B).

Of course, "bnd (B)" in this definition refers to the boundary *for a subset of a topological space*. [It could hardly be anything else. We do not have any "distances" now, to use for the definition of boundary for a metric space.]

Again, we check that our definition agrees with that for metric spaces when both are applicable.

Theorem. Let X, d, be a metric space, $X, \mathcal{J}$ its underlying topological space, and $A \subset X$. Then A is disconnected in the metric space X, d if and only if A is disconnected in the topological space $X, \mathcal{J}$.

Proof: A is disconnected in the metric space if and only if there exists a subset B of X such that some point of A is in B, some point of A is not in B, and no point of A is in

$\text{bnd}_d(B)$. But, by the theorem on page 63, $\text{bnd}_d(B) = \text{bnd}_{\mathcal{J}}(B)$. So, A is disconnected in the metric space if and only if there exists a subset B of X such that some point of A is in B, some point of A is not in B, and no point of A is in $\text{bnd}_{\mathcal{J}}(B)$, i.e., if and only if A is disconnected in the topological space.

Again, the theorem produces immediately examples, from those for metric spaces, of connected and disconnected subsets of various topological spaces. We do our usual two other examples.

Let $X, \mathcal{J}$ be a discrete topological space. Then for any $A \subset X$, bnd (A) is the empty set. Let $A \subset X$. When will A be disconnected? We must find a set B such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B). But the last will always hold, no matter what B is, for bnd (B) is the empty set. If A is the empty set, or A consists of just one point, then A will be connected. [For, under these circumstances, one could hardly find B such that some point of A is in B *and* some point of A is not in B.] Suppose, however, that A consists of two or more points. Let B be the set consisting of just one point, say p, of A. Then some point of A is in B (namely, p), some point of A is not in B (since A contains two or more points, while B consists of just p), and no point of A is in bnd (B) (since bnd (B) is empty). So, such an A is disconnected. Thus, in a discrete topological space, the empty subset and a subset consisting of just one point are both connected, while any other subset is disconnected.

Let $X, \mathcal{J}$ be an indiscrete topological space. Then bnd (B), for $B \subset X$, is the empty set if B is empty or $B = X$, and bnd $(B) = X$ otherwise. Now let $A \subset X$. When will A be disconnected? When will one be able to find B such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B)? We claim that one will never find such a B (no matter what A is). Indeed, since some point of A is supposed to be in B, B cannot be the empty set; since some point of A is supposed to not be in B, B cannot be X itself. But if B is neither of these, then bnd $(B) = X$. How can it happen, then, that no point of A is in bnd (B)? Only if A is the empty set. But finally, if A is the empty set, then it could hardly be that some point of A is in B. Thus, no matter what A is, we shall never be able to find a suitable set B in this topological space. In short, *every* subset of this topological space is connected. [This includes even, for example, an A consisting of just two points of X. This set, in this topological space, is connected!]

Finally, we recover the old theorem about connected sets.

Theorem. Let $X, \mathcal{J}$ be a topological space, A and B connected subsets of X, and p a point of both A and B. Then $A \cup B$ is connected.

Proof: Suppose, for contradiction, that some point of $A \cup B$ (say, u in A) is in C, some point of $A \cup B$ (say, v in B) is not in C, and no point of $A \cup B$ is in bnd (C), for some $C \subset X$. Then no point of A and no point of B is in bnd (C). But point u in A is in C and no point of A is in bnd (C), and so, since A is connected, every point of A must be in C. In particular, point p must be in C. But now point p in B is in C, point v in B is not in C, and no point of B is in bnd (C), which contradicts the fact that B is connected.

Again, the proof is just the old proof (page 23). Again, an "intuitively reasonable" theorem is recovered within the context of topological spaces.

Finally, we remark that all our old examples for the plane, and for metric spaces, become also examples for topological spaces, since the plane is a metric space and metric spaces give rise to topological spaces. In particular, all the things we found to be false before (provided those "things" use only interior, boundary, etc.) remain false for topological spaces.

I hope that some sense of the beauty of the definition of a topological space comes through all this. We put so little in (just give a set X and some subsets), and yet so much seems to come out. We shall see more of this shortly.

11. An Example: Connected Sets with Boundary Attached

All of the theorems we have dealt with so far have been relatively simple ones. This was necessary in order that complications inherent in the theorems themselves not obscure the transitions from the plane to metric spaces to topological spaces. We here give one example of a more complicated theorem. It will also serve to illustrate the idea that the notion of a topological space indeed succeeds in capturing a wide range of intuitive ideas.

Let A be a set in the plane. Then, intuitively, the boundary of A is "right up against A"; it does not contain "parts which are well-separated from A". These intuitive observations suggest the following. Suppose that A is connected. Let us now "attach to A its boundary", i.e., consider the set $A \cup \text{bnd}\,(A)$: Since bnd (A) would seem to be "right up against A", one might expect that $A \cup \text{bnd}\,(A)$ will also be connected. Each point of bnd (A) is in such "intimate contact" with A that it is hard to see how the attachment of such points to A could make a connected set A become disconnected.

All this suggests

Theorem. Let X, $\mathcal{J}$ be a topological space, and A a connected subset of X. Then $A \cup \text{bnd}\,(A)$ is connected.

Our next job, as topologists, would be to try to think of a proof. One might try first the idea of showing that bnd (A) must be connected and that A and bnd (A) have a point in common. Then, one could use the theorem that the union of two connected sets having a point in common is connected. Unfortunately, this will not work. Consider, for example, A the usual disk in the plane. Then bnd (A) (the "rim" of the disk) will *not* have points in common with A. Furthermore, we have seen that it is not true that the boundary of a connected set is connected. To come up with a proof that works, one might go back and look at the intuitive ideas which suggested the theorem in the first place. Suppose for a moment that, with A connected $A \cup \text{bnd}\,(A)$ could be "separated into two separate pieces". Why might one expect this to be impossible? Since A is to be connected, if such a separation were possible, then all of A would presumably have to be in "one of the pieces" – for otherwise we would have separated A into two pieces. Thus, one of our "pieces" would have to contain all of A (and possibly also some boundary points), while the other would have to contain *only* boundary points. But now, since these boundary points are not "well separated from A" (boundary points being as they are), it is hard to see how our two pieces could be "well separated", as required in the definition of disconnected. These ideas, written out more formally, generate the proof.

Proof: Suppose, for contradiction, that $A \cup \text{bnd}\,(A)$ were disconnected, so there is a subset C of X such that some point, p, of $A \cup \text{bnd}\,(A)$ is in C, some point, q, of $A \cup \text{bnd}\,(A)$ is not in C, and no point of $A \cup \text{bnd}\,(A)$ is in bnd (C). If p is in A, set $p' = p$. If p is not in A, then, since p is in $A \cup \text{bnd}\,(A)$, p must be in bnd (A). But p is in C and not in bnd (C), and so p must be in int (C). But int (C) is open, and so is an open neighborhood of p. Since p is in bnd (A), and int (C) is an open neighborhood of p, there exists a point, designated p', in A and int (C). But int $(C) \subset C$, and so p' is in A and C. Thus, in either case (whether p is in A or not), we obtain a point p' in A and in C. Similarly for q. If q is in A, set $q' = q$. If q is not in A, then q is in bnd (A). But q is in neither C nor bnd (C), and so must be in int (C^C). So, int (C^C) is an open neighborhood of q, and q is in bnd (A), and so there is a point, designated q', in int (C^C) and A. But int $(C^C) \subset C^C$, and so q' is in A and not C. Thus, in either case, we obtain a point q' in A and not C.

We now have a set C such that some point (namely, p') of A is in C, some point (namely q') of A is not in C, and no point of A is in bnd (C) (since no point of $A \cup$ bnd

(A) is in bnd (C)). But this contradicts the fact that A is connected.

This theorem is perhaps a bit more typical than the others of a nontrivial theorem in topology. What is perhaps surprising about it is that, while there is so little in the definition of a topological space, such results are actually true.

12. Compactness

We have generalized "interior", "boundary", and "connected" from metric spaces to topological spaces. We now turn to the generalization of "bounded".

One sees immediately that "bounded" is going to be more difficult. Consider, for example the definition of "interior" for a metric space. It reads "The interior of A is the set of all points p such that, for some positive number $\epsilon \ldots$", where "$\ldots$" refers only to the points within distance ϵ of p. So, the analogous definition for topological spaces becomes "The interior of A is the set of all points p such that, for some open neighborhood of $p, \ldots$". That is, we may just mechanically replace "the locus of points within ϵ of p" by "open neighborhood of p". But consider, by contrast, the definition of a bounded subset of a metric space: "Set A is said to be bounded if, for every positive ϵ, there exists a finite set of points of X, $p_1, \ldots, p_n$, such that every point of A is within ϵ of at least one of these points." The critical difference here is that one has to know the value of ϵ *before* one knows the finite set of points, $p_1, \ldots, p_n$. Indeed, as we have seen, one has to select one's finite set of points carefully, depending on what ϵ is. What we are demanding for bounded is that every point of A be in the locus of points within ϵ of p_1, or in the locus of points within ϵ of p_2, or in the locus $\ldots$ It is these locuses which we would like to replace by open neighborhoods. But how is this to be done, since we are to know the ϵ (i.e., the "sizes" of these open neighborhoods) prior to knowing which points they are to be open neighborhoods of? Thus, suppose one tries to write out an analogous definition of "bounded" for topological spaces: "Set A is said to be bounded if for every open neighborhood, there exists a finite set of points of X, $p_1, \ldots, p_n$, such that every point of A is in $\ldots$". How does one finish the sentence? Every point of A is in what? Our open neighborhood is just sitting there: We do not have open neighborhoods of $p_1, \ldots, p_n$. [If we did, then of course, one would demand that every point of A be in at least one of these open neighborhoods.] Furthermore, we say "if for every open neighborhood $\ldots$". Open neighborhood of what point? [Surely not of the p_i's, for we do not have access to the p_i's at this point in the definition.]

The key to finding something analogous to "bounded", but for topological spaces, is, as it turns out, to first reformulate somewhat the definition for metric spaces. This reformulation will be somewhat smoother using the following, easy, definition.

Definition. A collection of sets is said to *cover* set A if A is a subset of their union.

In other words, in order that a collection of sets cover A, it is necessary that *every* point of A be in at least one of those sets. [That is, it is essentially the same meaning as in "cover the wall with paint".]

Let X, d be a metric space, and $A \subset X$. It is easy to insert the word "cover" into the definition of boundedness of A. "Set A is said to be bounded if, for every positive ϵ, there exists a finite set of points of X, $p_1, \ldots, p_n$, such that the n sets, B_1 (the set of all points within ϵ of p_1), B_2, (the set of all points within ϵ of p_2), $\ldots, B_n$ (the set of all points within ϵ of p_n) cover A." [This just says the same thing as before, for demanding that the B's cover A is the same as demanding that every point of A is within ϵ of at least one of the p_i's.]

Of course, the mere insertion of the word "cover" does not immediately solve our problem. The next step is the one that contains the idea. Let X, d be a metric space, and $A \subset X$. Let a positive number ϵ be given. For p any point of X, let us write B_p for the set of all points q of X with $d(p, q) \leq \epsilon$. In the plane, for example, B_p would be a disk, of radius ϵ, centered at the point p. Now let us consider for a moment *all* subsets of X of the form B_p for p *any* point of X. In the plane, for example, this would be an enormous collection of disks, of radius ϵ, but with one for each point of the plane, centered at that point. Here, then, is a collection of subsets of X. This collection of sets certainly covers

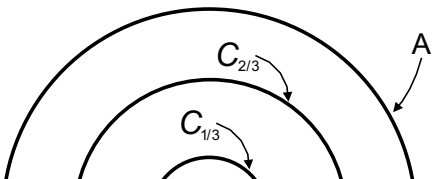
A. [Indeed, it even covers X. Given *any* point p of X, it is in at least one of the B's, namely, in B_p. Of course, this p may also be in other B's, but one is enough in order to cover.] So, we so far have $A \subset X$, we have positive ϵ given, and we have found a collection of sets, the B_p's, which cover A. Nothing has been said yet about "bounded". We now ask: How can one express, in terms of the B_p's, the statement that A is bounded? But this is easy. The statement is that a *finite* number of the B_p's will cover A.
[Indeed, "The finite number of B's, $B_{p_1}, B_{p_2}, \ldots, B_{p_n}$, cover A." means the same thing as "Every point of A is within distance ϵ of at least one of the points $p_1, p_2, \ldots, p_n$." Thus, "*Some* finite number of B's cover A." is the same as "*There exists* a finite set of points, $p_1, \ldots, p_n$, such that every point of A is within ϵ of at least one of them."]

So far, then, we have the following. Let X, be a metric space, and $A \subset X$. Then "A is bounded" means the same thing as "For any positive ϵ, a finite number of the B_p's (constructed as above *using the given* ϵ) cover A." Now for the final step in the translation. The B_p's are, of course, unsuitable for topology, for they are defined using the distance, while we shall not have distances in a topological space. By what should they be replaced? But they are practically self-interior. [They would have been self-interior, had we defined them as "all points q with $d(p, q) < \epsilon$" instead of "$\leq \epsilon$". But this difference, as we have seen, is not significant in these matters, for one can always replace ϵ by $1/2\epsilon$.] Thus, the plan is to replace "the B_p's" (for metric spaces) by "open sets" (for topological spaces). In carrying out the replacement, however, we would wish to retain, in the topological case, as much of the flavor of the B_p's in the metric case as we can. But there seems to be only one property of the B_p's that we say without using distance, namely that the B_p's cover A. So, we decide to replace "the B_p's" by a "collection of open sets which cover A". By what shall we replace the part about "For any positive ϵ"? Of course, the only role of ϵ is to tell us what B_p's we are to choose. thus, "for any positive ϵ" could be replaced by "for any collection of B_p's". [We emphasize that none of these replacements and translations are at all obvious. In practice, one would of course have to try a vast number of things, to see what works and what does not. One could easily obtain three or four generalizations – or none.]

In any case, these remarks suggest the following.
Definition. Let X, $\mathcal{J}$ be a topological space, and $A \subset X$. Then A is said to be *compact* if, for any collection of open sets which covers A, a finite number of these sets cover A.
It is conventional, in topology, to use the word "compact" rather than "bounded" – indeed, as we shall see, the two words have a somewhat different meaning. Thus, to show that a subset of a topological space is compact, one has to show that, no matter what collection of open sets one is given which covers A, some finite number of those will suffice to cover A. To show A not compact, one must find a collection of open sets which covers A, such that no finite number will do.

Some examples and elementary results will illustrate the definition.
Theorem. Let X, d be a metric space, X, $\mathcal{J}$ its underlying topological space, and A a compact subset of X in this topological space. Then A is bounded in the metric space.
Proof: Let a positive ϵ be given. for p any point of X, denote by B_p the set of all points q with $d(p, q) < \epsilon$. Then, by the theorem on page 58, each B_p is open in X, $\mathcal{J}$. But the B_p's cover A, and so, by compactness, a finite number $B_{p_1}, \ldots, B_{p_n}$, cover A. That is, every point of A is within ϵ of at least one of the finite number of points, $p_1, \ldots, p_n$. Thus, compact sets in the underlying topological space are always bounded in the metric space. Compare, for instance, with the theorem on page 60. (You get the same answer whether you take the interior in the metric space or in the underlying topological space.) This might suggest that the converse of this theorem should also be true: That any bounded set in the metric space must be compact in the underlying topological space.

Unfortunately, this converse is not true. Let X, d be the usual plane, and A the disk consisting of all (x, y) with $x^2 + y^2 < 1$. Then, as we



have seen, A is bounded in the metric space. Is A compact in the underlying topological space? Is it true that, given any collection of open sets which covers A, a finite number will do? For a any positive number less than 1, denote by C_a the set of points (x, y) with $x^2 + y^2 < a^2$, so C_a is "the disk of radius a centered at the origin". All of these C_a's are of course open in the underlying topological space. Clearly, they cover A. [Indeed, give any point p of A, so $d(p, \text{origin}) < 1$, choose a such that $d(p, \text{origin}) < \text{a} < 1$. Then p is in C_a.]

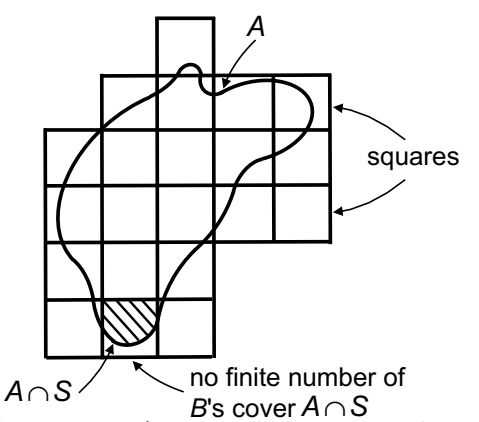
So, if A is to be compact, some finite number of these open sets had better cover A. But none will do. Indeed, suppose it was claimed that the finite number $C_{a_1}, \ldots, C_{a_n}$ cover A. Each of $a_1, \ldots, a_n$ must, of course, be a positive number less than one. Let a be the largest of these numbers, so a is less than one. Now choose point q with $a < d(q, \text{origin}) < 1$. Then this q will be in A, but will not be in any of our finite collection of C's. Thus, this disk is not compact.

Thus, at least for the plane, "compact" is stronger than just "bounded". However, it turns out that the disk is not compact for what is really a rather minor reason: We did not "attach to A its boundary". For sets "with boundary attached", in the plane, compactness means the same thing as boundedness. A simple, precise way to say "with boundary attached" is to demand that A^C be self-interior. Indeed, if A^C is self-interior, so $A^C = \text{int}(A^C)$, then since, by the second theorem on page 65, no point is in both bnd (A) and int (A^C), no point can be in both bnd (A) and A^C. That is, every point of bnd (A) must be in A. That is, A must include its boundary. We have

Theorem. Let X, d be the usual plane, and A a bounded subset of X with A^C self-interior. Then, in the underlying topological space, A is compact.

Unfortunately, a full proof of this requires some facts about real numbers which, while not terribly difficult, would require a rather lengthy explanation. So, we merely sketch the idea. Suppose, for contradiction, that A were not compact, so we have some collection of self-interior sets, B's, which cover A, but such that no finite number cover A. Cover A by a "grid" consisting of non-overlapping squares, each one unit by one unit. Since A is bounded, only a finite number of such squares will be needed.

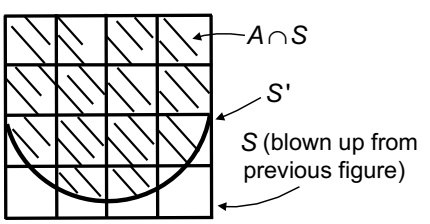
We now ask, for each such square S: Does a finite number of the B's cover $A \cap S$? The answer could be yes or no. However, the answer could hardly be "yes" for *every* one of our squares, for, if a finite number of B's covered the "part of A in S" for *every* square S, then, since there are only a finite number of squares, a finite number of the B's would cover A. So, there are must be at least one square S such that no finite number of B's cover $A \cap S$, say, the square indicated.



Now consider just this one square, and subdivide it into smaller squares, as shown. We now ask, for each of these smaller squares, S': Does a finite number of B's cover $A \cap S'$? Again, the answer cannot be "yes" for *every one* of the smaller squares, for if this were the case then a finite number of B's would have to cover $A \cap S$. So, choose one of the smaller squares, S', such that no finite number of B's covers $A \cap S'$. Now subdivide this S', into still smaller squares, ask the same question for each of these, and

find the still smaller square for each the answer is "no". Continuing in this way, we find a succession of smaller and smaller squares, each inside its predecessor. Let p be the point which is common to all of these squares. [It is for the existence of such a point of the plane that we need a property of the numbers.] We now ask the crucial question: Where is p, in A or not in A? Either answer will get us into trouble. Suppose p were not in A, so p is in A^C. But A^C is self-interior, so p is in int (A^C). That is, there is a positive number ϵ such that no point within ϵ of p is in A. But now one of our "succession of smaller and smaller squares, converging down on p" must lie entirely within this locus of points within ϵ of p. Call this square C, so C is a subset of A^C, i.e., C has no points in common with A. But recall how this C was to have been chosen (from a subdivision of a larger square): It was to be the one such that no finite number of the B's covers $A \cap C$. But this is a contradiction, for $A \cap C$ is the empty set, and certainly a finite number of the B's (such as, for example, any one of the B's) covers the empty set. Thus, "p not in A" leads to a contradiction. Suppose, then, that p were in A. But the B's are supposed to cover A, so one of the B's, say B_1, must contain the point p. Each B is self-interior. So, there is a positive number ϵ such that every point within ϵ of p is in B_1. But again one of our "succession of smaller and smaller squares, converging down to p" must lie entirely within this locus of points within ϵ of p. Call this square C, so C is a subset of B_1. But again C was to have been chosen, from a subdivision of a larger square, as the one such that no finite number of B's covers $A \cap C$. This is again a contradiction, for some finite number of the B's *does* cover $A \cap C$ in this case, namely the finite number consisting of only B_1. (For C is a subset of B_1, so B_1 covers C, so certainly B_1 covers $A \cap C$.) So, "p in A" also leads to a contradiction. Thus, we began by assuming that we had a collection of self-interior sets which cover A, but such that no finite number would do, and we obtain a contradiction. Hence, for any collection of self-interior sets which cover A, some finite number must cover A. That is, A must be compact.

Thus, for the plane, "compact" is not quite the same thing as "bounded", but it is quite close. One only has to deal with sets "having their boundaries attached", and the two notions agree. We give some more examples.



Theorem. Let $X, \mathcal{J}$ be a topological space, and A a finite subset of X. Then A is compact.

Proof: Let the points of A be $p_1, \ldots, p_n$. Consider a collection of open sets which covers A. Then one, say B_1, must contain p_1; and one, say B_2, must contain p_2; and so on to B_n, containing p_n. But now $B_1, B_2, \ldots, B_n$ is a finite number of sets in our collection which covers A.

Recall that every finite subset of a metric space is bounded. Again, compact is "similar" to bounded.

Let $X, \mathcal{J}$ be indiscrete, and $A \subset X$. Then A is always compact. Suppose first that A is empty. Then *any* set covers A. So, given any collection of open sets which covers A, a finite number (say, any one) will do. Suppose that A is not empty. Consider a collection of open sets which covers A. Since $X, \mathcal{J}$ is indiscrete, the only open sets are X and the empty set. But, since A is to be covered by these sets, at least one of them had better be X (for a lot of empty sets won't cover a not-empty A). But now a finite number of our open sets will do, namely just one of them, the "X".

Let $X, \mathcal{J}$ be discrete, and $A \subset X$. If A is finite, then, by the theorem above, A is compact. Suppose, then, that A is not finite. Then, we claim, A is not compact either. Indeed for each point of A, consider the subset of X consisting of just that point. These sets certainly cover A, i.e., every point of A is in one of them. Since our topological space is discrete, every subset of X is open, so these sets are open. But no finite number of these sets could cover A, for each set only contains one point, so a finite number could only cover a finite set, while A is not finite.

The union of two bounded sets in a metric space is bounded. Similarly,

Theorem. Let X, $\mathcal{J}$ be a topological space, and A and B compact subsets of X. Then $A \cup B$ is compact.
Proof: Let there be given a collection of open sets which covers $A \cup B$. Then this collection covers A, and so, since A is compact, some finite number, $C_1 \ldots, C_n$, covers A. Also, this collection covers B, and so, since B is compact, some finite number, $C_1{}' \ldots, C_m{}'$, covers B. But now the finite number $C_1 \cdots, C_n, C_1{}' \ldots, C_m{}'$, covers $A \cup B$.

In a metric space, any subset of a bounded set is bounded. This is false for compact sets in a topological space. For example, let B be the set of all (x, y) in the plane with $x^2 + y^2 \leq 1$, the "disk with boundary attached". Then, by the theorem on page 71, B is compact in the underlying topological space. Let $A \subset B$ consists of all (x, y) with $x^2 + y^2 < 1$. Then A is a subset of a compact set, but A is, as we have seen, not compact. Again, the problem has to do with "attachment of boundaries". What is true, however, is that subset A of compact B, provided A "has its boundary attached" is always compact.
Theorem. Let X, $\mathcal{J}$ be a topological space, B a compact subset of X, and $A \subset B$ with A^C open. Then A is compact.
Proof: Let there be given a collection of open sets which covers A. Include in this collection the open set A^C. Then, since the original collection covers A, and A^C covers A^C, the resulting collection covers X. In particular, it covers B. Since B is compact, a finite number of sets from this collection covers B, say $C_1, \ldots, C_n$ and A^C, where the C's are from the original collection. Since $A \subset B$, it must be that $C_1, \ldots, C_n$, A^C covers A. But, since no point of A is in A^C, if $C_1, \ldots, C_n$, A^C covers A then $C_1, \ldots, C_n$ must cover A.

Thus, a finite number of sets from the original collection covers A.
This is a tricky proof. One begins with a collection, the C's, of open sets which covers A. One tosses in an extra open set, A^C. The resulting collection covers B, and so, by compactness of B, a finite number will do. Of course, in this "finite collection" there may be included the extra set we threw in, namely A^C. Since $A \subset B$, this finite collection also covers A. But A^C can now be left out again, for it is not going to help any in covering A, since it has no points in common with A. Thus, we find a finite number of C's to cover A.

To summarize, "bounded" undergoes a curious transition from metric spaces to topological spaces. As the definition of "bounded" stands in metric spaces, it looks as though it will be impossible to generalize. One does obtain a generalization of sorts, to compactness, but it is not quite the same thing. *Very* roughly speaking, compact sets are like bounded sets which include their boundaries. But, compactness at least has much of the flavor of boundedness. In particular, whenever, in an argument, one would like to "think boundedness", one instead "talks compactness". There are enough properties in common so that one usually gets by. What is nice about compactness (aside, of course, from the most remarkable thing: that there should even *exist anything* analogous to "bounded" in topological spaces) is that the definition is so simple, and the proofs are so simple. Indeed, the proofs of the previous two theorems are perhaps simpler than the corresponding proofs for boundedness in metric spaces. It turns out that compactness is an extremely important and useful idea in topology – much more so than, for example, connectedness. Again, an intuitive idea which could easily have been lost in the transition to topology in fact survives.

13. Continuous Mappings of Topological Spaces

One of the central ideas in topology is that of a continuous mapping. In this section, we define continuous mappings, give some examples, and obtain some elementary properties.

Recall, from Sect, 5, our treatment of continuous curves in the plane. A curve was a rule which assigned, to each number t a point $\gamma(t)$ of the plane. We then considered continuous curves, those which "do not skip". A curve was said to be continuous, roughly speaking, if nearby $t-$values were assigned, according to the rule γ, to nearby points of the plane. The notion of a continuous mapping of topological spaces is a direct generalization of these ideas. What is it that goes into a continuous curve? Well, there is the set of real numbers, the set of points of the plane, and the rule γ which assigns, to each point of the first set, a point of the second. The part about "continuous", however, uses more structure; "For every positive number ϵ there exists a positive number δ such that, whenever $|t-t_0| \leq \delta$, $d(\gamma(t), \gamma(t_0)) \leq \epsilon$." What is this $|t - t_0|$? It is essentially the "distance", in the set of numbers between the point t of that set and the point t_0. Indeed, as we have seen, the set of numbers with this distance forms a metric space. Similarly, $d(\gamma(t), \gamma(t_0))$ makes use of the distance between two points of the plane. In short, what really goes into the notion of a continuous curve is i) the *metric space* of real numbers (with distance $|t-t_0|$), ii) the *metric space* the plane (with usual geometrical distance), and iii) a rule γ which assigns, to each point of the first metric space, a point of the second metric space. But the generalization, in topology, of a metric space is a topological space. Thus, the ingredients of what we are looking for should be i) a *topological space*, ii) a second *topological space*, and iii) a rule which assigns, to each point of the first topological space, a point of the second. Just as, for a *continuous* curve, one requires a certain property of the rule, so, for a continuous mapping of topological spaces, one will require a certain property.

These introductory remarks out of the way, we proceed to the definitions. We first wish to generalize "curve". This, from the observations above, is straightforward. Let X and X' be sets. Then a *mapping* from set X to set X' is a rule γ which assigns, to each point of the first set, X, a point of the second set, X'. For p a point of X, we write $\gamma(p)$ for the point of X' assigned by this rule to p. It is also convenient to use the symbol of $X \xrightarrow{\gamma} X'$ to mean "γ is a mapping from set X to set X'". We emphasize that, to specify a mapping, one must i) say what the set X is, ii) say what the set X' is and iii) give the rule γ. This rule must of course be unambiguous, it just specify a point of X' for *any* point of X, and it must yield *just one* point of X' in exchange for a point of X. [In fact, we shall be using mappings primarily when X and X' are the sets for topological spaces. But we do not need the open sets to say what a mapping is (any more than we needed the distances to say what a curve is), and so there is no harm in defining "mapping" for any sets X and X'.] It is ok if a mapping have the feature that some points of X' are not assigned from *any* point of X, or if two different points of X are sent, by the mapping, to the *same* point of X'. As long as γ specifies unambiguously a point of X' in exchange for a point of X, and as long as this specification is given for any point of X, this γ is a mapping.

"Mapping" is perhaps the second most important idea in all of mathematics. [It would be difficult to unseat "set" from first place.] We give some examples of mappings.

A curve (page 25) is just a mapping from the set of real numbers (X) to the set of points of the plane (X').

What is usually called a "function of one variable" in elementary mathematics is a rule which assigns to each real number, x, another real number, written $f(x)$. That is, this is just a mapping from X (the set of real numbers) to X' (also the set of real numbers).

Let X be any set, and let $X \xrightarrow{\gamma} X$ be the rule which assigns, to point p of X, the point p of X. [That is $\gamma(p) = p$.] This rule unambiguously gives a point of the second set (X) in exchange for a point of the first set (also X), namely "keep the same point". This is a mapping.

Let X be any set, and X' the plane. Let $X \xrightarrow{\gamma} X'$ be the rule which assigns, to any point p of X, the origin of X' (the plane), i.e., $\gamma(p) =$ origin. This is a mapping. [Here, of course, all points of X get sent by the rule to the same point of X', and so most points of X' (namely, all but the origin) do not come from any point of X.]

Let X, d be any metric space, and fix once and for all a point p of X. Let X' be the set of numbers. Let $X \xrightarrow{\gamma} X'$ be the following rule: Given point q of X, set $\gamma(q) = d)p, q)$. This indeed assigns a point of X' (i.e., a number) to each point, q, of X. So, we have a mapping. Here, for example, $\gamma(p) = 0$, by the first condition for a metric space.

Let Y, $\mathcal{J}$ be any topological space. Let X be the set of all subsets of Y. [Thus, a "point" of the set X is a subset of Y.] Let X' be a set consisting of just two elements, which we may denote "yes" and "no". Now let $X \xrightarrow{\gamma} X'$ be the following rule: Given a point p of X (that is, given a subset of Y), let $\gamma(p)$ be the point "yes" of X' if this subset of Y is open, and "no" if it is not open. Here is a rule which assigns to each point of X (i.e., subset of Y), a point of X' (i.e., the point "yes" or "no"). This is a mapping.

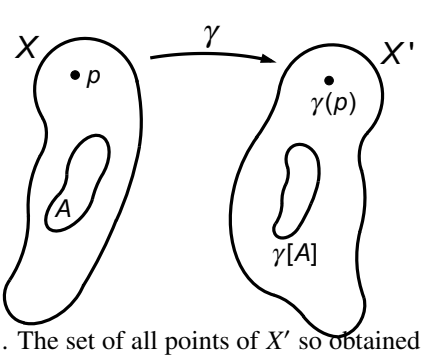
Let Y, $\mathcal{J}$ be any topological space. Let X be the set of all subsets of Y. Let $X \xrightarrow{\gamma} X$ be the rule which assigns, to any point p of X, the point bnd (p) of X. [In more detail, "point" p is a subset of topological space Y, $\mathcal{J}$. We can take the boundary, of this subset, and again obtain a subset of Y. That is, we again obtain a "point" of the set X. This point is $\gamma(p)$.] This is a mapping.

Let X be the plane, and X' be the real numbers. Let $X \xrightarrow{\gamma} X'$ assign, to point (x, y) of X, the number $x^3 - 3xy$. This is a mapping. We shall later need two easy definitions with mappings.

Definition. Let X and X' be sets, and $X \xrightarrow{\gamma} X'$. Let $A \subset X$. Then the *image* of A under γ, written $\gamma[A]$, is the set of all points of X' of the form $\gamma(p)$, with p in A.

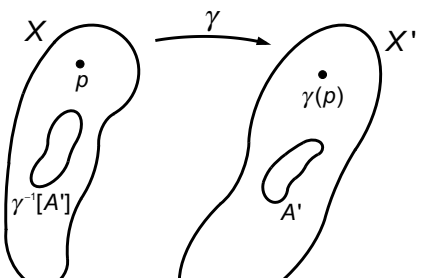
Definition. Let X and X' be sets, and $X \xrightarrow{\gamma} X'$. Let $A' \subset X'$. Then the *inverse image* of A' under γ, written $\gamma^{-1}[A']$, is the set of all points p of X with $\gamma(p)$ in the set A'.

The first definition is illustrated in the figure on the right. We have set X and X', and rule γ which assigns, to each point of X, a point of X'. further we are given some subset A *of* X. Now γ assigns, to *each* point of X a point of X'. In particular, we may consider the points of A (which of course are also in X, since $A \subset X$).



Given any such point, p, in A, we can apply our rule, and obtain a point, $\gamma(p)$, of X'. Let us now do this – apply γ – for *every* point of the set A. The result will be some points of X'. The set of all points of X' so obtained is the subset $\gamma[A]$ of X'. Thus, *one only takes the image of a subset of X, and the result,* $\gamma[A]$*, is always a subset of* X'. [Essentially, whereas the mapping starts out just sending each *point* of X to a *point* of X', consideration of the notion of "image" forces the mapping also to send each *subset* of X to a *subset* of X'.]

The second definition is illustrated in the second figure on the right above. Again, we have sets X and X', and mapping γ. Now, however, we consider subset A' *of* X'. Now γ assigns, to each point of X, a point of X'. Some of the resulting "points of X'" may happen to wind up in the subset A', and some may not.



We now consider all points *of* X which the rule γ *does* send into the subset A' of X'. This subset of X is the inverse image, $\gamma^{-1}[A']$. Thus, *one only takes the inverse image of a subset of* X'*, and the result,* $\gamma^{-1}[A']$*, is always a subset of* X. It is just the reverse of the "image". [It is easy to remember all this from the terminology. The image sends subsets of X to subsets of X', just as the original mapping sends points of X to points of X'. The inverse image is the reverse: It sends subsets of X' to subsets of X. We give some examples.

Let X be the set of numbers, X' the plane, and $X \xrightarrow{\gamma} X'$ a curve. Then $\gamma[X]$, the image X under this mapping, is the set of all points of the plane of the form $\gamma(t)$ for some number t. This is just "the set of all points of the plane through which the curve passes", i.e., the "points marked when we draw the curve".

Let X and X' both be the set of real numbers, and $X \xrightarrow{\gamma} X'$ the function of one real variable given by $\gamma(x) = x^2$. [Note: indeed a mapping.] Let A be the subset of X consisting of all numbers between 2 and 5. Then $\gamma[A[$, the image of A, is the set of all numbers (points of X') of the form $\gamma(x)$ for x between 2 and 5, i.e., the set of all numbers obtained by squaring a number between 2 and 5, i.e., the set of all numbers between 4 and 25. Next, let A' be the subset of X' consisting of all numbers between 4 and 25. Then $\gamma^{-1}[A']$ is the set of all numbers (points of X) such that $\gamma(x)$ is between 4 and 25. Hence, $\gamma^{-1}[A']$ consists of all numbers between 2 and 5, *and also* all numbers between (-2) and (-5) (for these also have squares between 4 and 25.

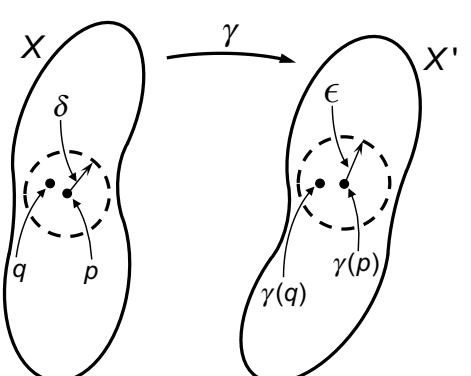
Let X be any set, and $X \xrightarrow{\gamma} X$ the identity mapping, with $\gamma(p) = p$ for every p in X. Then, for A a subset of X, $\gamma[A] = A$ (for the set of all points of the form $\gamma(p)$, with p in A, is just A, since $\gamma(p) = p$), and $\gamma^{-1}[A] = A$ (for the set of all points p with $\gamma(p)$ in A is just A).

Let X be any set, X' the plane, and $X \xrightarrow{\gamma} X'$ be the mapping with $\gamma(p) =$ origin, for any point p of X. Let A be any subset of X not the empty set. Then $\gamma[A]$ (set of points of X', the plane, one gets by applying γ to points of A) is the subset of X' consisting of just the origin. [Of course, γ [empty set = empty set.] Let A' be a subset of X' not including the origin. Then $\gamma^{-1}[A']$ (all points of X with $\gamma(p)$ in A' is the empty set (for $\gamma(p)$ is always the origin, while the origin is not in A'. So, $\gamma(p)$ is *never* in A', for any p.) For A' a subset of the plane X' including the origin, then $\gamma^{-1}[A']$ is all of X (for every point p of X has $\gamma(p)$ in A', since $\gamma(p)$ is the origin, and that is in A').

For the first example of two pages ago, γ^{-1}[yes] is just $\mathcal{J}$, the subset of X consisting of all open subsets of Y.

We can of course consider in particular mappings of topological spaces. That is: Let $X, \mathcal{J}$ and $X', \mathcal{J}$ and $X', \mathcal{J}'$ be topological spaces, and $X \xrightarrow{\gamma} X'$. Our next task is to define a continuous mapping of topological spaces. The plan, of course, is to use as guides the definition of a continuous curve, the observation that what was really involved in that definition was a *metric space* of numbers and really a *metric space* of the plane, and the fact that every metric space gives rise to its underlying topological space. What we want to do, in short, is define a continuous mapping of topological spaces in a way so that the result is reminiscent of the definition of a continuous curve.

So, let X, d and X', d' be metric spaces, and $X \xrightarrow{\gamma} X'$. [To be more concrete, one could let X, d be the metric space of numbers, with $d(t, t_0) = |t - t_0|$, and X', d' the plane with geometrical distance.] Fix a point p of X, and suppose that γ is continuous at p. That is, suppose that, for any positive number ϵ, there exists a positive number δ such that, when-

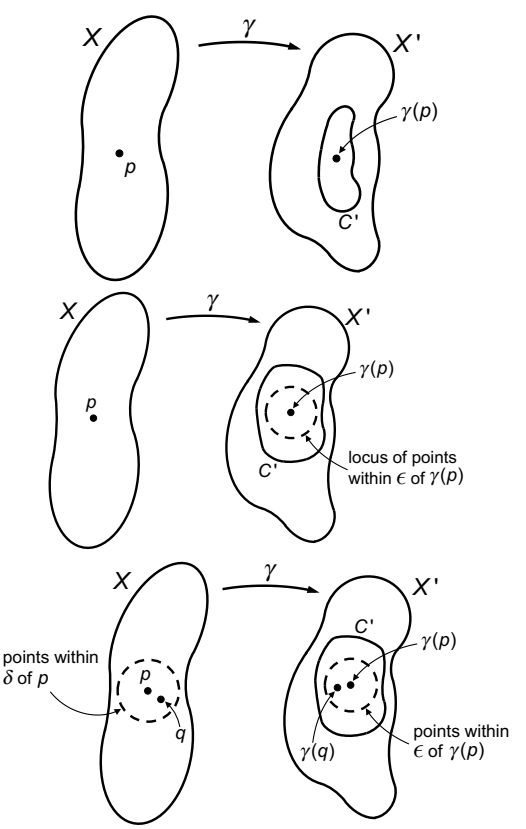


ever $d(p,q) \leq \delta$ (where q, of course, is in X), $d'(\gamma(p),\gamma(q)) \leq \epsilon$. [Note that this last makes sense: $\gamma(p)$ and $\gamma(q)$ are points of X', and d' gives distances between points of X'. But "$d'(p,q)$", for example, makes no sense, for d' does not give distances between points of X.] The plan is to eliminate all mention of "distance", in favor of "self-interior sets" in our metric spaces. The resulting statement, using just "self-interior sets", will then suggest a definition in topological spaces, replacing "self-interior" by "open".

So, we had better start by considering a self-interior set somewhere, i.e., in either X, d or in X', d'. Choosing one in X, d as it turns out, does not get one very far. [Of course, in practice, if one were trying to invent "continuous mapping of topological spaces" for the first time, one would have to try everything in sight. It might take one a full week to come to the realization that starting with a self-interior set in X, d is not very promising!] In any case, we decide to consider self-interior sets in X'. Since we are dealing with continuity at the point p of X, our self-interior set in X', d' had better have something to do with p. [We cannot, of course, just demand that p be in this self-interior set, for p is in X, while our self-interior set is in X'.] The next best thing would be to choose a self-interior set C' in X' such that (p) (a point *of* X') is in C'.

So, we so far have metric spaces X, d and X', d', point p of X, mapping $X \xrightarrow{\gamma} X'$ continuous at p (as above), and self-interior $C' \subset X'$ with $\gamma(p)$ in C'. Now, since C' is self-interior, we have $C' = \text{int}\,(C')$. So, since $\gamma(p)$ is in C', it is also in int (C'). That is to say, there exists a positive number ϵ such that every point q' of X' with $d'(\gamma(p), q') \leq \epsilon$ is in C'.

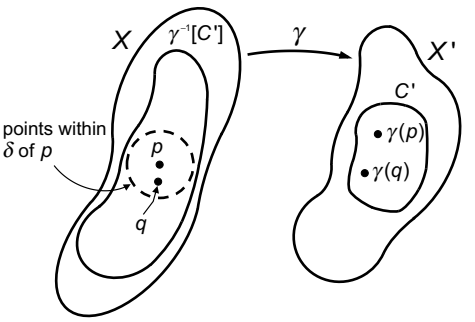
Now things are beginning to look promising. Recall the definition of continuous (on the left of the figure on the previous page): For any positive ϵ, there exists a positive δ such that, whenever $d(p,q) \leq$ *delta*, $d'(\gamma(p),\gamma(q) \leq \epsilon$. This must be true for *every* positive ϵ – and in particular for the ϵ we just obtained above (the one so that every point of X' within ϵ of $\gamma(p)$ is in C'). Thus, by continuity: There exists a positive number δ such that, for every point q of X with $d(p,q) \leq \delta$, $\gamma(q)$ is in C'.



Great! So far, by introducing a self-interior set in X', we have managed to eliminate all mention of the number ϵ. Next, we must eliminate mention of the number δ – and of course it must have something to do with self-interior sets in X'. How shall we do this? The key is to use the idea of inverse-image.

Just for fun, let us consider the inverse image of C', $\gamma^{-1}[C']$, i.e., the set of all points q of X with $\gamma(q)$ in C'. What can we say about this $\gamma^{-1}[C']$? Well, first of all, p is certainly in $\gamma^{-1}[C']$, for $\gamma(p)$ is in C' (for that is how we choose C' in the first place). What else must be in $\gamma^{-1}[C']$? The last sentence of the previous paragraph includes: "There exists a positive number δ such that, for every point q of X with $d(p,q) \leq \delta$, $\gamma(q)$ is in C'." In other words, using the definition of $\gamma^{-1}[C']$, "There exists a positive number δ such that every point q

with $d(p,q) \leq$ *delta* is in $\gamma^{-1}[C']$." But look at this last statement: It just says that p is in int $(\gamma^{-1}[C'])$. We have now eliminated mention of the number δ!



So far, then, we have the following. We have metric spaces X, d and X', d', point p of X, mapping $X \xrightarrow{\gamma} X'$ continuous at p, and self-interior subset C' of X' with $\gamma(p)$ a point of C'. We conclude under these circumstances that p must be in the interior of $\gamma^{-1}[C']$. This conclusion, of course, just expresses the idea that our mapping γ (of metric spaces) is continuous at p. But it has been cleverly arranged so that it uses words such as "self-interior", "inverse image", and "interior", and not the word "distance".

For the final step, we have to work on the point p. Let us now suppose that $X \xrightarrow{\gamma} X'$ is continuous, not only at the single point p, but even at *every* point of X. [Just as a continuous curve requires continuity at every point.] Let us keep fixed the self-interior set C' and consider changing our choice of the point p of X. We wish to invoke our conclusion above, "p must be in the interior of $\gamma^{-1}[C']$". For which p's will this conclusion follow? Well, all the conditions were listed above: "... metric spaces X, d and X', d', point p of X, mapping $X \xrightarrow{\gamma} X'$ continuous at p, and self-interior subset C' of X' with $\gamma(p)$ a point of C'." What of all this involves point p? First, the mapping γ must be continuous at p: But this is ok, for we have just assumed the mapping continuous at *every* point of X. Second, the point p must be such that $\gamma(p)$ is in C'. So, this is the only condition on p. That is to say, we have that p must be in the interior of $\gamma^{-1}[C']$ *whenever* $\gamma(p)$ is in C'. But "$\gamma(p)$ is in C'" is the same as "p is in $\gamma^{-1}[C']$". So, we have that "every point p of $\gamma^{-1}[C']$ must be in the interior of $\gamma^{-1}[C']$". But look again at this last quote. [Of course, every point of the interior of $\gamma^{-1}[C']$ is in $\gamma^{-1}[C']$, by an earlier theorem.] It just says that $\gamma^{-1}[C']$ is self-interior! Thus, not only do we eliminate the number δ, but everything reduces finally to talking about self-interior sets.

What we have shown, then, is the following. Let X, d and X', d' be metric spaces, and $X \xrightarrow{\gamma} X'$. Then, if γ is continuous at every point of X [that is, if for every point p of X and every positive number ϵ there exists a positive number δ such that, whenever $d(p,q) \leq \delta$, $d(\gamma(p), \gamma(q)) \leq \epsilon$ – what a mess!], it follows that the inverse image under γ of every self-interior subset of X' is a self-interior subset of X. By just going back trough the discussion above, one convinces oneself that the converse is also true, i.e., that if inverse images of self-interior sets are self-interior, then γ is continuous (i.e., all that $\epsilon - -\delta$ business holds). In short, we have succeeded in expressing the very messy idea of continuity for mappings of metric spaces just in terms of self-interior sets. The resulting expression is extremely simple: no "you choose this and I choose that", no complicated business about getting the order of choices right. All one has to do is take inverse images of self-interior sets in X', and check to see if the resulting subsets of X are also self-interior.

The last step is very easy. We just generalize from metric spaces to topological spaces by replacing "self-interior" by "open". All this suggests, then, the following.

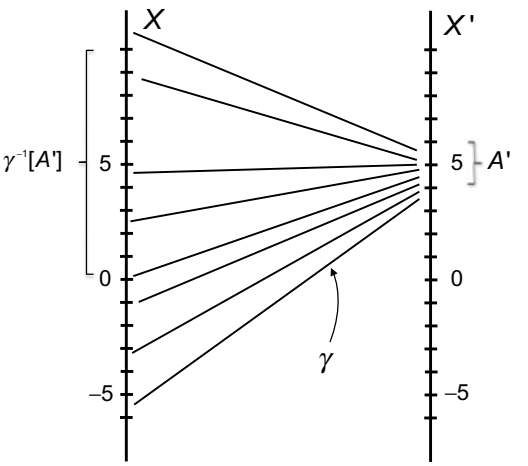
Definition. Let $X, \mathcal{J}$ and $X', \mathcal{J}'$ be topological spaces, and $X \xrightarrow{\gamma} X'$. Then γ is said to be *continuous* if the inverse images under γ, of all open subsets of X' are open subsets of X. The hope, then, is that this definition captures the idea that the mapping sends, "nearby" points of X, to "nearby" points of X' i.e., that the point $\gamma(p)$ of X' does not "skip" or "jump around" in X' when p moves just a little bit in X. As we all shall see, it does capture this idea.

We give some examples.

Let $X, \mathcal{J}$ be any topological space, and $X \xrightarrow{\gamma} X$ the identity mapping. Then, for any

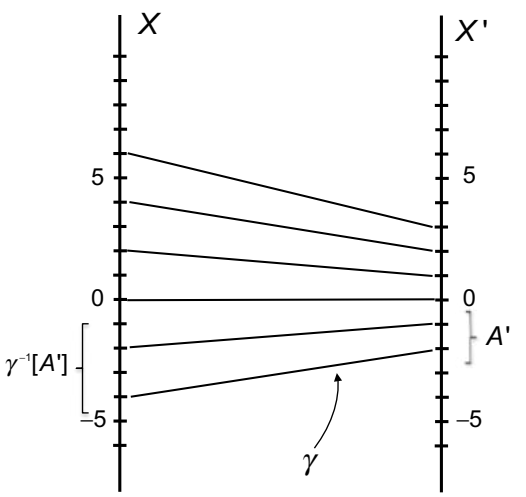
subset A of X (the one of the right), $\gamma^{-1}[A] = A$ (a subset of the X on the left). Obviously, inverse images of open sets are open. This mapping of topological spaces is continuous. [Intuitively, $\gamma(p)$ move in the X on the right just as p moves in the X on the left. There is no "jumping around".]

Let each of X, $\mathcal{J}$ and X', $\mathcal{J}'$ be the topological space of real numbers (i.e., the underlying topological space of the metric space of numbers, with distance $|t - t_0|$.) Let $X \xrightarrow{\gamma} X'$ be the following: Given a point t of X (i.e., a real number), let $\gamma(t)$ be 7 (a point of X') if $t < 0$, and 5 if $t \geq 0$. This is a mapping. [Intuitively, $\gamma(t)$ "jumps" from the point 7 of X' to the point 5 as t increases through 0. One would not expect it to be continuous.] This mapping is not continuous.

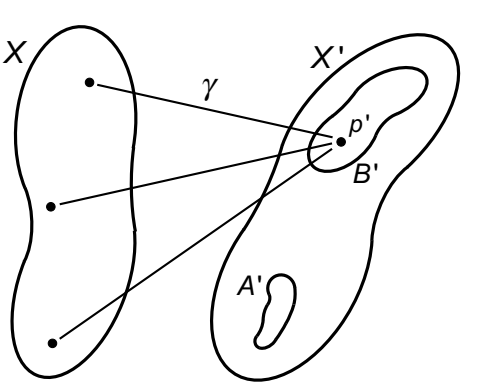


Let A' be the subset of X' consisting of all numbers x with $41/2 < x, 51/2$. This set is self-interior in the metric space, and so is open in the underlying topological space. Now $\gamma^{-1}[A']$ is the set of all numbers t (in X) with $41/2 < \gamma(t) < 51/2$. But, from γ is, this is just the set of all numbers t in (X) with $t \geq 0$. But this $\gamma^{-1}[A']$ is not a self-interior subset of X (for the number 0 is in this set, but not in its interior). That is, $\gamma^{-1}[A']$ is not open in the topological space X, $\mathcal{J}$. So, we have found an open set in X', $\mathcal{J}'$ whose inverse image is not open in X, $\mathcal{J}$. So, this mapping is not continuous.

Let X, $\mathcal{J}$ and X', $\mathcal{J}'$ be as above, but let $X \xrightarrow{\gamma} X'$ be the mapping with $\gamma(t) = 1/2t$. [That is, this is the mapping "divide by two". Intuitively, division by two should take nearby numbers to nearby numbers. So, we expect this mapping to be continuous.] This mapping is continuous. Let A' be any open subset in X', $\mathcal{J}'$. Then $\gamma^{-1}[A']$ is just A' "blown up by a factor of two", i.e., all numbers of the form $2x$, for x in A'. This $\gamma^{-1}[A']$ must be a self-interior subset of the metric space. [Indeed, given a point, $2x$, of $\gamma^{-1}[A']$, find the ϵ which shows that x is in int (A'). Use 2ϵ to show that $2x$ is in int $(\gamma^{-1}[A'])$.] Thus, the inverse image of any open set in X', $\mathcal{J}'$ is open in X, $\mathcal{J}$.



Let X, $\mathcal{J}$ and X', $\mathcal{J}'$ be topological spaces. Fix once and for all a point p' of X'. Let $X \xrightarrow{\gamma} X'$ be the mapping with $\gamma(q) = p'$, for any q in X. [Intuitively, "all of X is sent to the single point p' of X'". There could hardly be any jumping around of $\gamma(q)$, since it is always the point p'. So, one expects continuous.] This mapping is continuous. Let A' be an open subset of X', with p' not in A'. Then



$\gamma^{-1}[A']$ is the empty subset of X – an open
set, by the first condition for a topological
space. Let B' be an open subset of X',with p' in B'. Then $\gamma^{-1}[B']$ is all of X – an open set, by the first condition for a topological space. Thus, inverse images of open sets are open.

Let X, $\mathcal{J}$ be a discrete topological space, X', $\mathcal{J}'$ a topological space, and $X \xrightarrow{\gamma} X'$. Then this mapping is always continuous. Let A' be an open set in X', $\mathcal{J}'$. Then $\gamma^{-1}[A']$ is certainly *some* subset of X. But X, $\mathcal{J}$ is discrete, and so this is an open subset of X (for every subset of X is open in this topology). So, inverse images of open sets are open. [Intuitively, think of X, $\mathcal{J}$ as coming from the metric space, with $d(p, q) = 0$ if $p = q$, and 1 otherwise. Continuity is to ask whether "small changes" in the point p result in "large changes in $\gamma(p)$. But the points "p" are to be in this metric space. The only way one can get a "small change" from p (at least, if this "small change" is motion by a distance less than 1) is just to remain at p. Any other "change" of the point p moves one a whole distance of 1. So, of one only moves a "little bit" from p, i.e., if one stays at p, then $\gamma(p)$ will not change at all (because p could not change). So, indeed, "small changes" in the point p *do* result in small changes in $\gamma(p)$. So, one would expect continuous.]

Let X, $\mathcal{J}$ be a topological space, X', $\mathcal{J}'$ an indiscrete topological space, and $X \xrightarrow{\gamma} X'$. Then this mapping is always continuous. Let A' be an open subset of X'. Then, since X', $\mathcal{J}'$ is indiscrete, A' must be either the empty subset of X', or X' itself. For A' empty, $\gamma^{-1}[A']$ is the empty subset of X; for $A' = X'$, $\gamma^{-1}[A']$ is all of X. But, by the first condition for a topological space, both of these are open subsets of X. [Intuitively, any change of point in an indiscrete topological space is "small": The open sets are so insensitive to the points that can hardly tell one point from another. Since the open sets can not really tell the points apart, they regard any change of point as "small". So, when p chances a little bit, $\gamma(p)$ always changes a little bit, for the latter is a "change" in X', $\mathcal{J}'$, where everything is a "little bit". The indiscrete topological spaces are like "metric spaces" in which all distances are zero (although, of course, these are not metric spaces). If all distances are zero, every change of point is always by a "small distance".]

Every continuous curve defines a continuous mapping on the underlying topological spaces.

Let X, d be any metric space, and fix a point p of X. Let X' be the set of real numbers, and $X \xrightarrow{\gamma} X'$ the mapping with, for any point q of X, $\gamma(q) = d(p, q)$, a point of X' (i.e., a number). [Intuitively, the mapping is "evaluate the distance from p'". One would expect that "nearby points in X will have nearly equal distances from p". That is, one would expect the mapping to be continuous.] This mapping is continuous. We sketch the argument. Let A' be self-interior in the metric space of real numbers;: we must show that $\gamma^{-1}[A']$ is self-interior in X. Let q be any point of $\gamma^{-1}[A']$, i.e., let $\gamma(q) = a$ (a real number, namely $d(p, q)$) be in A'. We must show that q is in the interior of $\gamma^{-1}[A']$. That is, we must find a positive number δ such that every point of X within δ of q is in $\gamma^{-1}[A']$. That is, we must find positive δ such that every point r within δ of q has $\gamma(r)$ in A'. Since A' is self-interior (in X'), there is a positive number ϵ such that every number within ϵ of $\gamma(q) = a$ is in A'. Set $\delta = \epsilon$. Let r be any point within $\delta = \epsilon$ of q.

Then by the triangle inequality, $d(p, q) + d(q, r) \geq d(p, r)$, or $a + \epsilon \geq d(p, r)$; and $d(p, r) + d(r, q) \geq d(p, q)$, or $d(p, r) + \epsilon \geq a$, or $d(p, r) \geq a - \epsilon$. But these two together say that $d(p, r) \leq a + \epsilon$, and $d(p, r) \geq a - \epsilon$. That is, they say that $d(p, r)$ is within ϵ of a. That is, they say that $\gamma(r)$ is within ϵ of a. But every number within ϵ of a is in A' (for that is how we choose ϵ). So, $\gamma(r)$ must be in A'. Thus, we have found our positive number δ such that every point r within δ of q has $\gamma(r)$ in A' (namely, $\delta = \epsilon$). So, inverse images of self-interior sets are self-interior. So, in the underlying topological spaces, the mapping is continuous.

I hope these examples at least suggest that the definition of "continuous" is a reasonable candidate for saying what we want to say. One strong piece of evidence that

"topological space" is the right thing to be considering is that "continuous mapping" becomes so simple an idea in its framework. As the last example illustrates, a very simple definition can become quite complicated when it is actually applied.

14. Applications of Continuous Mappings

Many of the more interesting results in topology involve the interaction between continuous mappings on the one hand, and properties of various subsets of topological spaces on the other. In this section, we give some examples.

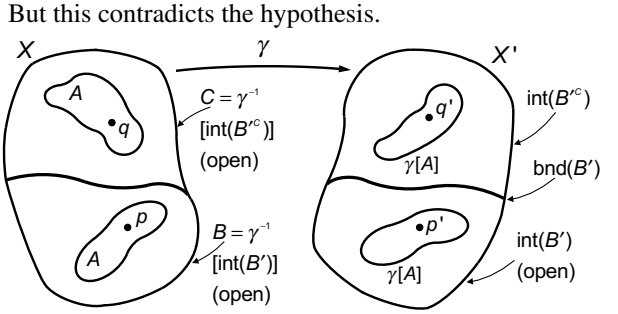
Let X, $\mathcal{J}$ and X', $\mathcal{J}'$ be topological spaces, and $X \xrightarrow{\gamma} X'$ a continuous mapping. Intuitively, γ "sends nearby points of X to nearby points of X'", i.e., it "sends X to X' without tearing". Now let A be a connected subset of X. Intuitively, A "consists of just one piece in X". In terms of these intuitive characterizations, one might expect that "since A is just one piece, and since no tearing is allowed under the mapping, the corresponding image set in X' must also consist of just one piece" – for, after all, this would seem to describe what we mean by "no tearing". These remarks suggest a result in topology.

Theorem. Let X, $\mathcal{J}$ and X', $\mathcal{J}'$ be topological spaces, $X \xrightarrow{\gamma} X'$ a continuous mapping, and A a connected subset of X. Then $\gamma[A]$ is a connected subset of X'.

Proof: Suppose, for contradiction, that $\gamma[A]$ were disconnected, so there is a subset B', of X' such that some point, p', of $\gamma[A]$ is in B', some point q', of $\gamma[A]$ is in B', some point, q' of $\gamma[A]$ is not in B', and no point of $\gamma[A]$ is in bnd B'). Set $B = \gamma^{-1}[\text{int}(\text{B}')]$, a subset of X.

We shall show that i) some point of A is in B, ii) some point of A is not in B, and iii) no point of A is in bnd (B). i) Since p' is in $\gamma[A]$, there is a point p of A with $p' = \gamma(p)$. Since p' is in B' and not bnd (B'), p' is in int (B'). So, since $\gamma(p) = p'$, p is in $\gamma^{-1}[\text{int}(\text{B}')] = \text{B}$. So, p is in A and B. ii) Since q' is in $\gamma[A]$, there is a point q of A with $q' = \gamma(q)$. But $\gamma(q)$ is not in B', and so not in int (B'), and therefore q cannot be in $\gamma^{-1}[\text{int}(B')] = B$. So, q is in A and not in B. iii) Suppose, for contradiction, that there were a point r in A and bnd (B), Then $\gamma(r)$ is in $\gamma[A]$, and so, since no point of $\gamma[A]$ can be in bnd (B'), $\gamma(r)$ must be in either int (B') or int (B'^C). Were $\gamma(r)$ in int (B'), then r would be in $\gamma^{-1}[\text{int}(B')] = B$. But int (B') is open in X', $\mathcal{J}'$, and so, by continuity of γ, B is open in X, $\mathcal{J}$, and so $B =$ int (B). So, since r is in B, r is in int (B) – and hence r cannot be in bnd (B). Were $\gamma(r)$ in int (B'^C), r would be in $C = \gamma^{-1}[\text{int}(B'^C)]$. But int (B'^C) is open in X', $\mathcal{J}'$, and so, by continuity of γ, C is open in X, $\mathcal{J}$. So, C is an open neighborhood of r. But r is in bnd (B), and so, since C is an open neighborhood of r, there must be a point u in both C and B. Since u is in C, $\gamma(u)$ must be in int (B'^C), while, since u is in B, $\gamma(u)$ must be in int (B'). But no point of X' can be in both int (B'^C) and int (B'). We conclude, then, that the existence of a point r in both A and bnd (B) leads to a contradiction. Hence, there is no point of A in bnd (B).

We supposed that $\gamma[A]$ were disconnected, and found a set B such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B), That is, we found A disconnected, But this contradicts the hypothesis.

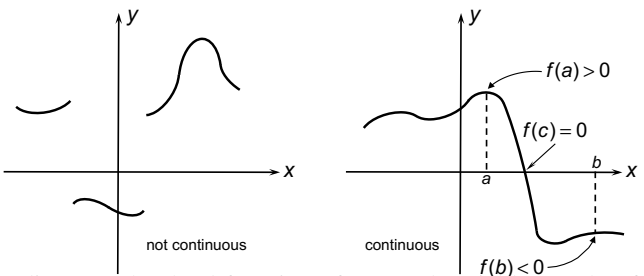


This is of course quite a complicated proof. We have connected $A \subset X$, and we suppose that $\gamma[A]$ is disconnected, in order to obtain a contradiction. We consider the set

B' which "disconnects" $\gamma[A]$. The idea is to use this B' to show that A is disconnected, i.e., to arrive at a contradiction. This is done by taking a lot of inverse images. The "two"pieces" into which $\gamma[A]$ is separated are represented by int (B), and int (B'^C). [All that remains is bnd (B') – but no points of $\gamma[A]$ are in there.] These are open sets in X', and so, by continuity, their inverse images, called B and C above, are open sets in X. These are the two open sets which will "separate A into two pieces". That B actually has the three properties necessary to "disconnect" A (some point of A in B, some point of A not in B, and no point of A in bnd (B)) is shown by using for each, the corresponding property for B' "disconnecting" $\gamma[A]$. We essentially "take the inverse image of each property". The first two are easy, while the third, iii) above, is a bit tricky – using, for example, the definition of boundary.

The thrust of the theorem is that, again, the actual definitions of "continuous" and "connected" seem to reflect our intuition.

We give two examples of more "concrete" applications of this theorem.



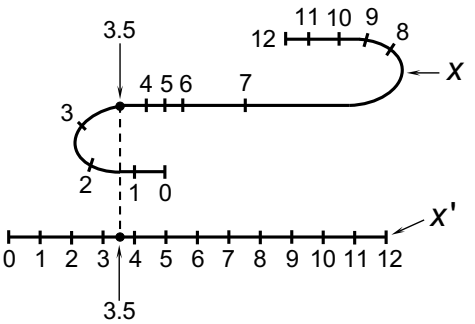
Consider ordinary real-valued function of one real variable. Such a function can be represented by its graph (of y as a function of x, say). Continuity for such a function means, roughly speaking, that the graph is "one continuous line, with no jumps". It seems "obvious" that, if a *continuous* function somewhere becomes positive (i.e., $f(a)$ is positive for some number a) and somewhere becomes negative (i.e., $f(b)$ is negative for some number b), then that function somewhere is zero (i.e. $f(c) = 0$ for some number c). This is true.

Theorem. Let each of X, $\mathcal{J}$ and X', $\mathcal{J}'$ be the usual topological space of real numbers, and $X \xrightarrow{f} X'$ a continuous mapping. Let, for some a in X, $f(a) > 0$, and, for some b in X, $f(b) < 0$. Then, for some c in X, $f(c) = 0$.

Proof: For contradiction, suppose not. Let $A = X$, a connected subset of X. Then, by the previous theorem, $f[A]$ is a connected subset of X'. Let $B' \subset X'$ be the set of all positive numbers. Then some point, $f(a)$, of $f[A]$ is in B', and some point, $f(b)$, of $f[A]$ is not B'. But bnd (B') consists of just the number 0, while we have supposed that for no c does $f(c) = 0$. So, no point of $f[A]$ is in bnd (B'). We conclude, thus, that $f[A]$ is disconnected, a contradiction.

That is, we consider the connected subset X of X. Its image must be connected. But this image contains a positive number, $f(a)$, and a negative number, $f(b)$ – so how could this image be connected unless it also contains 0? [The B' makes this explicit.] But $f[A]$ containing zero means exactly that $f(c) = 0$ for some c in X.

The second application of our theorem is one of a family of results called fixed-point theorems in topology. Imagine two rubber rules (whose markings run, say, say, from zero to twelve inches). Leave one ruler alone on the table, and take the other ruler, stretch or bend it in any way one wants (but no tearing!), and lay it next to the first ruler. This is to be done such that no part of the second ruler extends beyond the first ruler. A typical example of



the result of this operation is shown in the figure. It turns out that, no matter how one stretches, bends, or folds (provided no spindling or mutilating) the second ruler, there will always exist some mark on the second ruler which winds up directly in the line with the corresponding mark on the first ruler! In the example above, for instance, the mark "3.5 inches" on the second ruler is in line with the mark "3.5 inches" on the first ruler. A few trials will convince one that this always works out this way (but it is not entirely obviously).

There is a theorem is topology which reflects this experiment. We replace the ruler left on the topological space. Let X' be the set of numbers x with $0 \leq x \leq 12$, and let the topology be the usual one. [That is, first make X' a metric space, with distance $d(x, x') = |x - x'|$, the usual distance between numbers. Then take the underlying topological space.] The ruler which is stretched and bent is represented by the same topological space, which we write X, $\mathcal{J}$. So, we so far have our rulers represented by topological spaces. The operation of "bending and stretching and laying next to" is just a thinly disguised mapping, $X \xrightarrow{\gamma} X'$, for this operation indeed assigns, to each point of the "bent and stretched" ruler, a point of the "left alone" ruler. That no tearing is allowed is expressed by demanding that this mapping be continuous. Finally, the conclusion we wish to draw is that some mark on the stretched ruler ends up directly in line with the corresponding mark on the stationary ruler. This is expressed by asking that $\gamma(a) = a$ for some point a of X (the stretched ruler). Thus, we have translated our "experiment" into mathematics. We have:

Theorem. Let X, $\mathcal{J}$ and X' and $\mathcal{J}'$ be the topological spaces above, and $X \xrightarrow{\gamma} X'$ continuous. Then, for some point a of X, $\gamma(a) = a$.

Proof: Suppose, for contradiction, that for no point a of X does $\gamma(a) = a$. Let X'' be the set consisting of just two points, p and q with the discrete topology. Let $X \xrightarrow{\mathcal{J}} X''$ be the following mapping: If $\gamma(a) > a$, set $\mathcal{J}(a) = p$; if $\gamma(a) < a$, set $\mathcal{J}(a) = q$. This is indeed a mapping, since, by supposition, $\gamma(a) = a$ for no a.

We show that $\mathcal{J}$ is continuous. It suffices to show that $\mathcal{J}^{-1}[p]$ and $\mathcal{J}'^{-1}[q]$, subsets of X, are open. To this end, let a be a point of $\mathcal{J}^{-1}[p]$, so $\gamma(a) = a' > a$. Set $\epsilon = 1/2(a' - a)$, a positive number. Denote by B' the set of points x' of X' with $|a' - x'| < \epsilon$, so B' is open in X', $\mathcal{J}'$. By continuity of γ, $\gamma^{-1}[B'] = B$, the set of points x of X with $|\gamma(a) - \gamma(x)| < \epsilon$, is open in X, $\mathcal{J}$. Let C be the set of points x of X with $|a = x| < \epsilon$. Then C is also open in X. By the second condition for a topological space, $B \cap C$ is also open in X – and clearly a is a point of $B \cap C$. So, $B \cap C$ is an open neighborhood of a in X. Let b be any point of $B \cap C$. Then b is in B, and so $\gamma(b)$ is within ϵ of $\gamma(a)$; and b is in C, and so b is within ϵ of a. But $\gamma(a) = a + 2$, by our choice of ϵ. So, we have $\gamma(b) > b$. That is, b is in $\mathcal{J}^{-1}[p]$. Thus, for any point, a, of $\mathcal{J}^{-1}[p]$, we have found an open neighborhood, $B \cap C$, of a, such that any point of $B \cap C$ is in $\mathcal{J}^{-1}[p]$. That is, we have shown that $\mathcal{J}^{-1}[p]$ is open in X, $\mathcal{J}$. Similarly, $\mathcal{J}^{-1}[q]$ is open. So, $\mathcal{J}$ is continuous.

Since X is open connected and $\mathcal{J}$ is continuous, $\mathcal{J}[X]$, by the theorem on page 82, is connected in X''. But X'', $\mathcal{J}''$ is discrete, and so, in order that it be connected, $\mathcal{J}[X]$ must consist of just the point p or just the point q. Now certainly $\gamma(0) > 0$, so $\mathcal{J}(0) = p$, so p is in $\mathcal{J}[X]$; and $\gamma(12) < 12$, so $\mathcal{J}(12) = q$, so q is in $\mathcal{J}[X]$. So, both p and q are in $\mathcal{J}[X[$: a contradiction.

The crucial part of the proof is to show that, if $\gamma(a) > a$, then "sufficiently nearby" a''s have also $\gamma(a') > a'$. This, of course, is what one would expect, from continuity of the mapping γ. The business about ϵ is just to pin down "how sufficiently nearby" a' must be. Once we have this, we have that $\mathcal{J}^{-1}[p]$, and similarly $\mathcal{J}^{-1}[q]$, are open – so that $\mathcal{J}$ is a continuous mapping. We just "invented" X'' and $\mathcal{J}$ in order to be able to use the theorem on page 82. We now do so, to conclude that *both* p and q cannot be in $\mathcal{J}[X]$ – i.e,. that either $\gamma(a) > a$ for all a, or $\gamma(a) < a$ for all a. Finally, we obtain a contradiction, since $\gamma(0) > 0$ and $\gamma(12) < 12$. Note that, if we "leave the end points off the rulers", i.e., consider only x with $0 < x < 12$, then the proof above fails, in the last step. In fact, it is

not difficult to find, for for "rulers without endpoints" a way to match up the two rulers so that no mark of X is aligned with the corresponding mark of X'.

We give a second example of an application of continuous mappings. Let X, $\mathcal{J}$ and X', $\mathcal{J}'$ be topological spaces, and $\xrightarrow{\gamma} X'$ continuous. Intuitively, γ "bends and stretches, but does not tear". Let A be a compact subset of X, so, intuitively, A is "well bounded in" in X. One might expect, then, that the image of A, $\gamma[A]$, must also be "well bounded in", for this could fail only if the action of γ is to "stretch A by an infinite amount, to make its image become unbounded", while a "stretching by an infinite amount" is practically a "tear". [This "argument", perhaps, is not a very convincing one. The problem seems to be that there is not a terribly good description, in non-technical words, of "compact".] In any case, such a result is true.

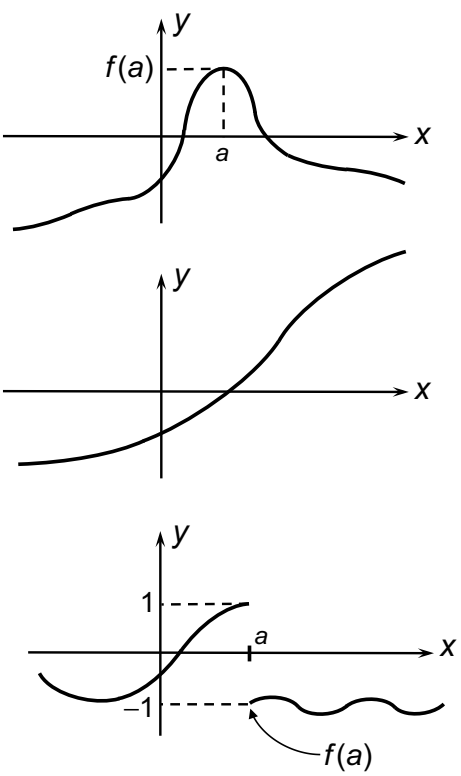
Theorem. Let X, $\mathcal{J}$ and X', $\mathcal{J}'$ be topological spaces, $X \xrightarrow{\gamma} X'$ continuous, and A a compact subset of X. Then $\gamma[A]$ is compact.

Proof: Suppose, for contradiction, that $\gamma[A]$ were not compact. Let the B''s be a collection of open sets which covers $\gamma[A]$, such that no finite number covers $\gamma[A]$. For each B' in this collection, set $B = \gamma^{-1}[B']$, so, by continuity of γ, each subset B of X is open. Foe any point p of A, $\gamma(p)$ is of course in $\gamma[A]$, and so $\gamma(p)$ is in one of the sets B' (since they cover $\gamma[A]$), and so p is in $B = \gamma^{-1}[B']$. Thus, every point of A is in one of the B's i.e., the B's cover A. But A is compact, and so a finite number of the B', say $B_1, B_2, \ldots, B_n$ cover A. We claim that, as a consequence, the corresponding sets in X', $B'_1, \ldots, B'_n$, cover $\gamma[A]$. Indeed, let p' be any point of $\gamma[A]$, so $p' = \gamma(p)$ for some point p of A. Since $B_1, \ldots, B_n$ cover A, p must be in one of these, say B_1. But p in $B_1 = \gamma^{-1}[B'_1]$ means that $p' = \gamma(p)$ is in B'_1. So, this p' is in one of $B'_1, \ldots, B'_n$, namely B_1.

We conclude, then, that a finite number of the B''s cover $\gamma[A]$ – which contradicts our choice of the B''s.

The argument is rather like that on page 82. We just use inverse images to "pull back" things from X' to X. Continuity, of course, plays a central role, for it ensures that we do not ruin open sets when we "pull them back to X".

We give one example of an application of this theorem. Consider again real-valued functions of one real variable. We say that such a function "assumes its maximum" if there is a number a such that $f(a)$ is "at least as large as f of anything ever gets to be". For example, the function illustrated in the first graph on the right assumes its maximum – for $f(a)$, as indicated, is at least as large as any $f(b)$. The second function illustrated does not assume its maximum: In fact, it does not have any maximum at all, for $f(x)$ "just keeps getting bigger as x gets bigger". The third function illustrated doe not "keep getting bigger and bigger" – but still it does not assume its maximum. The problem here is that, even though the function itself never exceeds the value one, the function is so arranged that $f(a)$ is actually -1: The function makes a "jump" at a, to avoid having $f(a) = 1$. [Of course, were $f(a)$ one, then the function would assume its maximum.] Under what conditions might one expect a func-



tion to assume its maximum? The "reason" why the function in the last example does not assume its maximum is that there is a "jump" just where it ought to be assuming its maximum. We can prevent this possibility by demanding that the function be continuous. For the other example, the problem is that the function "trails up to large values for $f(x)$ as x trails off to infinity". We can prevent this sort of thing by just considering the function for x's in some "finite region", say $0 \leq x \leq 1$. Then, of course, we shall only be able to show that the function "assumes its maximum compared with other values it assumes for x with $0 \leq x \leq 1$".

Are these safeguard sufficient to ensure that a function assume its maximum? The next theorem will show that they are. In the proof, we shall make use of two elementary facts about numbers: i) The set of all numbers x with $0 \leq x \leq 1$ is compact (virtually a consequence of the theorem on page 71), and ii) given any compact subset of the set of real numbers, there is an element of that subset which is larger than any other element of that subset.

Theorem. Let each of X, $\mathcal{J}$ and X', $\mathcal{J}'$ be the usual topological spaces of real numbers, and X' , $\mathcal{J}'$ be the usual topological spaces of real numbers, and $X \xrightarrow{\gamma} X'$ a continuous mapping. Let $A \subset X$ be all points x of X with $0 \leq x \leq 1$. Then there exists a point a of A such that $f(b) \leq f(a)$ for any point b of A.

Proof: By the theorem above, since A is a compact subset of X, $f[A]$ is a compact subset of X'. So, there exists a point a' of $f[A]$ is such that every point b' of $f[A]$ satisfies $b' \leq a'$. Since a' is in $f[A]$, there is a point a of A with $f(a) = a'$. Then, for any point b of A, $f(b)$ is in $f[A]$, and so $f(b) \leq f(a)$.

These, of course, are just a few selected examples. They do not begin to do justice to the enormous range and complexity of result in this subject. I hope, however, that they do illustrate the facts that topology does make contact with more "concrete" things, and that the original program – to build this subject entirely on the definition of a topological space – is at least a promising one.

15. Conclusion

Topology is one of the branches of modern mathematics. It is of course difficult to divide any subject – and mathematics in particular – into neat, clearly defined, "branches". An attempt to do so in the case of mathematics might yield two others: algebra and analysis. Algebra is essentially the study of "operations", of the results of manipulating elements of sets according to certain, prescribed, rules. The objects with which one deals (analogous to topological spaces in topology) are such things as groups and vector spaces. If one had to say in a sentence the difference between algebra and topology, it would perhaps be that in algebra one is concerned primarily with individual elements of sets; in topology, with subsets. Analysis, on the other hand, involves functions (usually, of real or complex variables), derivatives, integrals, differential equations, and so on. It is, for example, the area in which calculus largely lies. What is usually called "algebra" in high school, however, is perhaps 30% algebra, and 70% analysis. Of the three branches, algebra and topology are perhaps are the most self-contained. Analysis often drawn heavily on the other two – and in particular on topology. There are, of course, many areas of mathematics which lie between these branches (for example, Lie groups, which are almost exactly halfway between algebra and topology, with a dash of analysis), and others (such as axiomatic set theory) which are separate from them all.

Mathematics finds application to the other sciences, and so topology, as a part of mathematics, itself finds application. While it would be impossible to describe in detail such applications here, a few examples – all to physics – may illustrate this idea.

Consider a box containing a gas. Denote by X the set of points of the box (so, for example, we might regard X as the set of points, (x, y, z) of Euclidean three-dimensional space, with $0 \leq x \leq 1$ $0 \leq y \leq 1$, and $0 \leq z \leq 1$). At each point of the box, one can determine the temperature, T, of the gas. Of course, different points could have different temperatures: The gas could be cold at the bottom of the box, and warm at the top. Thus, "measurement of the temperature" could be represented as a rule which assigns, to each point of X (point of the box), a number (the temperature at that point). We thus acquire a mapping from X to X', the set of real numbers. One knows in physics that heat flows from regions of high temperature to regions of low temperature. Furthermore, the faster the temperature changes (say, per inch of motion in the box), the faster the heat flows from one region to the other. A consequence of this is that one cannot have the temperature "change suddenly" between two nearby points of the box, for the flow of heat would immediately cool the hot region and warm the cold region. That is to say, the temperature, as a function of position in the box, cannot "jump". This would be reflected in the mathematics as follows. We would regard the set X as a topological space (the underlying one, from geometrical distance in Euclidean space), and the set X' as a topological space (the usual one). Then, we would demand that our mapping be continuous. Thus, a physical idea, about the way temperature behaves, would be represented as a mathematical one, about continuity of a mapping. Then, for example, the theorem on page 82 would say that, if two temperatures are achieved in the box, then every temperature in between is also achieved, somewhere in box. The theorem on page 85 would say that there is some "hottest point" in the box.

A mechanical system in physics is initially described by a set X which represents "the set of all possible configurations of the system", and is called the configuration space of the system. For example, consider the system of a ball free to roll on a table. A point of X would then be represented by giving the location of the ball on the table, and the origination of the ball in space. [In this example, the configuration space would be five-dimensional – two dimensions to tell where the ball is on the table, and three to tell the orientation of the ball.] One has a physical sense of what it means for two configurations

to be "nearby". For example, two configurations, each having the ball at the same point of the table, but with one having the ball slightly rotated relative to the other, would be regarded as "nearby" – as would two configurations having the ball with the same orientation but at slightly different places on the table. This physical sense would be a reflected by regarding X as a topological space. Thus, as part of our physical ideas, the configuration space of a system is to be a topological space.

Now suppose that the table is tilted somewhat, so the ball, when released, will roll. Consider the experiment "begin with the ball at rest in some configuration, release the ball, and see what its configuration is five seconds later (after the ball has rolled a bit)". This is a rule which assigns, to each configuration (the initial one), some configuration (the final one). This would be represented, then, by a mapping from X to X. One aspect of the idea of dynamical stability in physics is that, if one changes very slightly the initial configuration of the ball, its ultimate configuration will not be changed very much, either. If this were not true, then the slightest little disturbance of the initial configuration would change dramatically what happens to the ball – something we just do not see happen in Nature. Thus, "dynamical stability" in physics would be reflected in our mathematical description by demanding that our mapping "with five seconds" be a continuous mapping of topological spaces.

But topology – as I hope has been made clear, – is not just a handmaiden of physics. It is an intellectual endeavor in its own right. It seeks to capture, within the framework which is mathematics, the notion of "closeness". The boundary of a set is the set of points which are "close to the set", and also "close to being out of that set"; a continuous mapping of topological spaces is one which preserves the "closeness" characterized by those spaces. This "captured closeness" – this notion of a topological space – has a sense of distance in it: Not the exact numerical values for distances, but rather enough of the shadow of distance to talk about what one wants to talk about.

Topology, at least for me, is a particularly pretty and aesthetic subject. Two aspects of the subject make it so. First, one puts so little into the definitions, yet so many tricky theorems come out. There just is not much to a topological space – just a set with some subsets satisfying a couple of simple properties. The various definitions – of connected, continuous mappings, and so on – are about as simple as definitions could be. But, on these few elements, one can build theorems which are very complicated indeed: for example, that the continuous image of a connected set is connected; that the attachment of the boundary to a connected set yields a connected set. These few theorems, of course, just scratch the surface. There are numerous theorems in this subject which are many times more complicated than these. [And, of course, we have presented only a small fraction of the definitions in topology.] The definitions build on each other, and intertwine, to produce a most appealing mosaic. The second remarkable thing about topology is that, not only does one have access to many complicated theorems, but furthermore these theorems on the whole even "make sense" intuitively. Surely there is no obvious reason why things work out this way. Look at the definition of a topological space. It certainly does not seem to scream out "Here is represented the essence of our ideas of 'closeness'." But that is exactly what is represented! How does it happen that such simple definitions, by means of complicated theorems which intertwine those definitions, turn out to mean something to our geometrical intuition? This is the sort of thing one usually wants in mathematics – and it is exactly what happens in topology.

But topology has also played a second role: It has served as an example for all of mathematics. Mathematics, as we have seen in this example, proceeds on two levels. On the one hand, there is the intuitive level, that in which one gets ideas, invents definitions which say what they are supposed to say, decides what statements are likely to be theorems and what are not, decides what line is likely to produce a proof of a theorem. The other level is that of precision. One proves theorems. Here on just uses the *statements* of the definition, the hypotheses of the theorems, and so on, not what those things are as ideas. One mechanically follows the rules, and produces proofs. Herein, at least for

me, lies much of the charm of mathematics. It is almost as though there are two entirely different subjects at work here. From each, one only dimly sees the other. But, despite all this, they end up saying virtually the same thing. The second level – the precise one – is of course the more difficult for the beginner. One is used to making great leaps, of getting to the answer, to using every thought one can generate. In constructing proofs, however, one must hold oneself back. One must simply follow the rules (which one has made for oneself!) and go where one is led.

So, when one begins mathematics, there is a hard part (writing proper proofs), and an easy part (getting ideas, guessing what is true and what false). But, as one goes on, the roles reverse. One soon becomes accustomed to writing out correct proofs. it becomes mechanical, once one has seen the idea of the proof. The hard part – the stuff of much research in mathematics – becomes thinking of new definitions which "say" new things, finding new and unexpected relationships between old definitions, and even finding counterexamples to show that "widely expected" relationships simply do not hold.

The style of mathematics is not well-adapted to everything. It would be a bit silly, for example, to begin psychology with "Definition. A *person* is a . . .", and later obtain "Theorem. There are exactly two sexes." For those things to which it *is* adapted, however, it delivers certain benefits. The definitions and theorems in mathematics "live forever". They are not to be changed by "subsequent evidence". [But, of course, there are styles in mathematics as in everything. Some definitions or theorems may later be regarded as uninteresting because of change of taste.] There is a kind of permanence in mathematics one does not see in other intellectual endeavors. Mathematics gives a new meaning to the word "true". But all this is at a price. Often, it is that one has to work very hard to show something which is "obvious". For example, "if one attaches the boundary to a connected set, the result is connected." seems clearly, under any reasonable interpretations of the words, "true". But we must, in mathematics, work comparatively hard for the proof. Sometimes all this work pays off. Something which seemed clearly to be true is just plain false. It usually means that we did not understand intuitively, in the fullest sense, the nature of the objects with which we were (or thought we were) dealing. More often, the work does not pay off so directly. It is just the price one pays for being able to say "true" as a mathematician.

Mathematics is not to everyone's taste, any more than any other intellectual endeavor. Why, if one does not care for it, should one know anything about it? I think that there may be something of value in mathematics for almost everyone. I do not mean in balancing the checkbook: That is hardly "mathematics". Mathematics forces one to a way of thinking very different from that in which one is accustomed. Perhaps a flexible mind is a viable end in itself. Stretching exercises, unfortunately, are not easy to come by. One might, for example, search for societies out of contact with Western civilization, join such a society for a few years, and come to understand how the minds of the members work. We are rapidly running out of these. Fortunately, mathematics, while perhaps not entirely "out of contact with western civilization", can serve. In short, I feel that mathematics can be useful because it is different in a very real sense – and different it certainly is.

Finally, I hope that I have managed to convey the fact that there is such a thing as creativity in mathematics – and such a thing as excitement.

Appendix

Phy Sci 112 February 17, 1978

Is the Curve Continuous?

Rules: This is a game for two players, Yes and and No. The two players begin by agreeing on a curve γ (i.e., a rule which assign, to each number t, a point of the plane), and also on a number t_0. The game itself consists of three moves: 1. No announces a positive number ϵ. 2. Yes announces a positive number δ. 3. No announces a number t such that $|t_0 - t| \leq \delta$. At the completion of the three moves, the players together determine (using the rule γ) the points $\gamma(t_0)$ and $\gamma(t)$ of the plane, and then the number $d(\gamma(t_0), \gamma(t))$. If $d(\gamma(t_0), \gamma(t)) \leq \epsilon$, then Yes wins the game. Otherwise, No wins.
Strategy: No will win at the end if it is not true that $d(\gamma(t_0), \gamma(t)) \leq \epsilon$. No wants, with his first move, to make this as unlikely as possible. He does so by choosing ϵ to be a small positive number. Yes will win at the end if it is true that $d(\gamma(t_0), \gamma(t)) \leq \epsilon$. Yes, of course, can do nothing about ϵ; but he does control δ in his second move. The role of δ is to restrict the range of No's choices of t in the third move. Yes wants to restrict No as much as possible, and so he chooses δ to be a small positive number. No, for his third move, already knows ϵ and δ. So, in order to win, he must find t with $|t_0 - t| \leq \delta$, but not $d(\gamma(t_0), \gamma(t)) \leq \epsilon$.
General Remarks: This is of course a game of pure skill: There is no luck. Either player can certainly blow the game by using poor strategy. [For example, No could lose a winning game in the third move by selecting a poor t; Yes could choose δ too large in the second step, and thus give No too many possibilities for the t he is to choose in the third move.] Suppose, however, that both players are experienced, and use the best possible strategy. Then, clearly, the winner is already predetermined by the choice of the curve γ and the number t_0. Problem: State what must be true of γ and t_0 in order that Yes, using prefect strategy, can always win the game. Answer: Your statement will be exactly the definition of a curve continuous at t_0.

Supplementary Problems Due: January 27, 1978

[Not to be turned in. Caution: These problems were "test rejects". In some cases, they were rejected because they were too ambiguous, or because they were too hard.]

1. Think of a set in the plane as "discrete" if every point of A is "separate from" the rest of A; no point of A is "intermingled with" the rest of A. Invent a definition of "discrete".

2. Give a formal proof that the boundary of a bounded set is bounded.

3. Is it true that, for any sets A and B in the plane, bnd $(A \cup B)$ is a subset of bnd $(A) \cup$ bnd (B)?

4. Let A, B, and C be connected sets in the plane, and let A and B have a point in common and B and C a point in common. Does it follow that the union of all three sets is connected?

5. Is it always true that int $(\text{bnd}\,(A)^C) = \text{bnd}\,(A)^C$?

6. Find sets A and B in the plane such that bnd $(A) = B$ and bnd $(B) = A$.

7. Is bnd (int $(A)) \subset$ bnd (A) for every set A in the plane?

8. One thinks of a set as disconnected if it "consists of two or more pieces". Try to find a definition of "one of those connected pieces of which the set consists". The definition would read: "For A a set in the plane, a subset of B of A is a *connected piece* of A if . . ."

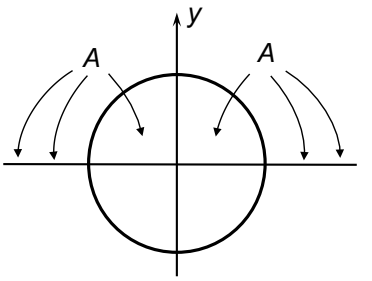
9. Let n be a positive integer, and let A be a set with exactly n points. For which values of n is A connected? For which values bounded? What is int (A)?

10. Let A be the set of all points (x, y) with $x^2 \neq y^2$. Find int (A) and bnd (A). Is A connected? bounded?

11. Find A such that neither A nor A^C is connected.

Phy Sc 112 **Test 1** January 30, 1978

1. (12 points) Let A be the set in the plane consisting of all points (x, y) with $x^+y^2 < 1$ or $y = 0$, i.e., a "disk with a line running through it". Indicate in a clear diagram the sets bnd (A) and int (A). Give the formulae for these sets. Is A bounded? Is A bounded? Is A connected? [Please answer each of these last two in a sentence or two, with a reason.]



2. (20 points) Please answer in a sentence or two, with a reason.

i) Is it true that, for A any bounded set in the plane, int (A) is bounded?

ii) Is it true that, for any two sets in the plane having no point in common, their union is disconnected?

iii) Is it true that, for any set A in the plane with int (A) connected, A must be connected?

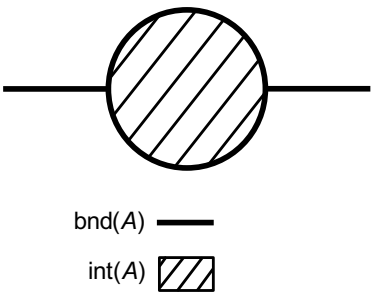
iv) Is it true that, for any set A in the plane, every point of A is in either int (A) or bnd (A)?

3. (10 points) Theorem. Let A be a set in the plane. Then no point of the plane is in both int (A) and bnd (A). [Give a formal proof.]

4. (10 points) Definition. A set A in the plane is said to be *closed* if, for every point p not in A, there exists a positive number ϵ such that no point q with $d(p, q) \leq \epsilon$ is in A. Is the usual disk (all (x, y) with $x^2 + y^2 < 1$) closed? Is the line (all (x, y) with $y = 1$ closed)? [Please answer in a sentence or two, with a reason.]

5. (8 points) Find a disconnected set A in the plane such that bnd (A) is the line by all (x, y) with $y = 1$.

Test 1 – Solutions



1. The sets bnd (A) and int (A) are as shown in the diagram. bnd (A) is the set of all points (x, y) with $x^2 + y^2 = 1$, or $y = 0$ and $x > 1$, or $y = 0$ and $x < -1$. int (A) is the set of all points (x, y) with $x^2 + y^2 < 1$. Set A is not bounded (and, indeed, the line itself is not even bounded): There is no point p of the plane and positive number c such that every point of A is within distance c of p. The set A is connected, because each of the line and disk by itself is connected, these sets have a point in common, and A is their union.

2. i) Yes. Int (A) is a subset of A, and every subset of bounded set is bounded. ii) No. Let A be be any set in the plane, and set $B = A^C$. Then A and B have no points in common, while their union, the entire plane, is connected. iii) No. Let, for example, A consists of the usual disk together with the point $(17, 17)$. Then A is certainly not connected, while int (A) is just the disk, which is connected. iv) Yes. Let p be a point of A. If for some positive ϵ every point within ϵ of p is in A, then p is in int (A). If this does not hold (i.e., if for every positive ϵ there is a point within ϵ of p and not in A), then, since of course there is for every positive ϵ a point, namely p, within ϵ of p and in A, p must be in bnd (A).

3. Proof: Let p be a point of int (A), so, for some positive number ϵ, every point within distance ϵ of p is in A. Then it cannot be the case that for every positive ϵ there is a point within ϵ of p and not in A. Hence, p cannot be in bnd (A).

4. The disk is not closed. The point $p = (1, 0)$ is not in A, and yet for every positive number ϵ there is a point within ϵ of p and in A. The line is closed. Given any point p not in A, let ϵ be one-half the perpendicular distance from p to the line, a positive number. Then it is indeed true that no point within distance ϵ of p is in A.

5. Let, for example, A be the complement of the line, i.e., the set of all points (x, y) with $y \neq 1$. Then A is disconnected, but its boundary is the line. [A second type of example is: Let A consist of all (x, y) with $y = 1$ and $x \neq 13$.]

Phy Sc 112 **Test 2** February 20, 1978

1.(6 points) Give an example of a metric space in which every bounded set is connected.

2. (16 points) For each of the statements below, tell whether it is true or false, and explain in a few sentences:

i) Let X, d be a metric space. Then int $(X) = X$.

ii) Let X, d be a metric space, and $A \subset X$. Then, if A and A^C are both bounded, X is bounded.

iii) Let X, d be a metric space, and $A \subset X$. Then if A and A^C are both connected, X is connected.

iv. Let X, d be *any* metric space with X the plane. Let A be the usual disk (all (x, y) with $x^2 + y^2 < 1$). Then int $(A) = A$.

3. (8 points) Consider the curve in the plane given by $\gamma(t) = (1, t)$ if $t < 1$, and $\gamma(t) = (1, 1)$ if $t \geq 1$. Is this curve continuous at $t = 1$? [If yes, tell how to choose δ given ϵ; if no, tell how to choose ϵ.]

4. (20 points) Let X, d be a metric space, and let p be a point of X. Let A be the subset of X consisting of all points q of X with $d(p, q) < 1/2$. Which of the following are *always* true (i.e., no matter what X, d, and p are) [Please answer in a sentence or two.]:

i) int $(A) = A$

ii) bnd (A) is the set of all points q with $d(p, q) = 1/2$

iii) A is bounded

iv) A is connected

5. (10 points) Let X, d be a metric space, and let A be a bounded subset of X. Then bnd (A) is bounded. [Give a formal proof.]

Test 2 – Solutions

1. Let X be the set consisting of just one point, p and let $d(p, p) = 0$. This is a metric space. The subset of X consisting of just p, and the empty subset of X, are clearly both connected. Thus, every subset of X is connected, and so certainly every bounded subset of X is connected.

2. i) True. Every point of int (X) is certainly in X. For the converse, let p be a point of X. Then every point of X within distance 231 of p is in X, and so p is in int (X). ii) True. X is the union of A and A^C, while the union of two bounded sets is bounded. iii) False. Let X consist of two points, p and q, and let $d(p, p) = d(q, q) = 0$, and $d(p, q) = 1$. This is a metric space. Let A consist of just p. Then A^C consists of just q, and so A and A^C are connected. But X is not connected. iv) False. Let X be the plane, let d' be the usual geometrical distance, and set $p = (0, 0)$ and $q = (2, 0)$. For r and s points of X neither p nor q, set $d(r, s) = d'(r, s)$, $d(p, r) = d(r, p) = d'(r, q)$, $d(q, r) = d(r, q) = d'(r, p)$ and $d(p, q) = d(q, p) = d'(p, q)$. This is a metric space! In it, int (A) is all points (x, y) with $0 < x^2 + y^2 < 1$, which is different from A.

3. This curve is continuous at $t = 1$. Given positive ϵ, any δ less than or equal to ϵ will do. Indeed, $d(\gamma(t), \gamma(1))$ is zero if $t \geq 1$, and is $d((1, t), (1, 1)) = 1 - t$ if $t < 1$. So, if $|1 - t| \leq \delta$, and $\delta \leq \epsilon$, then $d(\gamma(t), \gamma(1)) \leq \epsilon$.

4. i) Always true. By a theorem, int $(A) \subset A$. For the converse, let q be in A, say $d(p, q) = a < 1/2$. Choose positive ϵ such that $a + \epsilon < 1/2$. Then any point within ϵ of q is, by the triangle inequality, within $1/2$ of p, and so is in A. Since every point within ϵ of q is in A, q is in int (A). ii) Not always true. Let X be the plane, and let $d(p, q) = 0$ if $p = q$, and $d(p, q) = 1/2$ otherwise. Then A consists only of the point p, and bnd (A) is the empty set. But the set of all points q with $d(p, q) = 1/2$ is all of X except for the point p. So, bnd (A) is not the set of all q with $d(p, q) = 1/2$. iii) Not always true. Let X be the plane, and $d(p, q) = 0$ if $p = q$ and $d(p, q) = 1/4$ if $p \neq q$. Then $A = X$. But X itself is not bounded. iv) Not always true. The example of iii) in which A is certainly not connected will do.

5. Let positive ϵ be given. Then since A is bounded, there is a finite set of points of X, $p_1 \ldots p_n$, such that every point of A is within distance $\epsilon/2$ of at least one of them. Let p be any point of bnd (A). Then some point, say q, of A is within $\epsilon/2$ of p. But now, by triangle inequality, (since $d(p, q) \leq \epsilon/2$ and $d(q, p_i) \leq \epsilon/2$) we have $d(q, p_i) \leq \epsilon$. Thus, we have shown that every point of bnd (A) is within ϵ of at least one these p_i. So, since ϵ is arbitrary, bnd (A) is bounded.

Phy Sc 112 **Test 3** March 10, 1978

1. (20 points) Let X be the plane, and let $\mathcal{J}$ consist of the empty set, X itself, and the subset of X consisting of all (x, y) with $x > 0$. So, $X, \mathcal{J}$ is a topological space. Let $A \subset X$ consist of all (x, y) with $x^2 + y^2 < 1$.

i) Find int (A), and explain your answer.

ii) Find bnd (A), and explain your answer.

iii) Is A connected? Explain your answer.

iv) Let $X', \mathcal{J}'$ be the plane with the usual topology, and $X \xrightarrow{\gamma} X'$ the identity mapping. Is γ continuous? Explain.

2. (16 points) true or false? Please explain in a sentence or two.

i) The intersection of any two open neighborhood of point p in a topological space is an open neighborhood of p.

ii) The union of any two open neighborhoods of point p in a topological space is an open neighborhood of p.

iii) The image of an open set under a continuous mapping of topological spaces is an open neighborhood of p.

iv) The image of a disconnected set under a continuous mapping of topological spaces is a disconnected set.

3. (9 points) Theorem. Let $X, \mathcal{J}$ be a topological space, and A and B open subsets of X having no point in common. Let $A \subset$ bnd (B). Then A is empty. [Give a formal proof.]

4. (15 points) A topological space $X, \mathcal{J}$ is said to be *locally connected* if every point p of X has a connected open neighborhood. Is every discrete topological space locally connected? Every indiscrete topological space? The plane with the usual topology? [Please explain why.]

Note: "The plane with the usual topology" is the underlying topological space of the plane with usual geometrical distance.

Problem Set 1 Due: January 16, 1978

1. Draw a picture of the set of points (x, y) in the plane satisfying $xy = 1$.

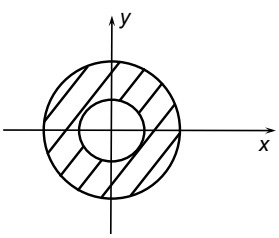
2. Find a simples characterization of each of the following two sets in the plane: i) those (x, y) satisfying $x^2 = y^2 \leq 0$; ii) those (x, y) such that $x \geq 0$ or $y \geq 0$ or $x + y \leq 0$

3. Which of the following describe sets in the plane? i) all (x, y) such that, for some $y, x = y$; ii) Start with the three points $(0, 0)$, $(0, 1)$ and $(1, 0)$. Then include any point lying on a line joining any two of these. Then include any point lying on a line joining any two of the resulting points. etc. iii) all (x, y) such that $x + y + z = 3$.

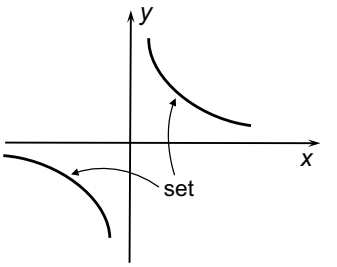
4. We have given examples of various statements which do, and various which do not, define sets in the plane. Find the best example you can of a statement which is borderline between the two.

5. Consider the following two properties of sets in the plane: i) "is a one-dimensional set", and ii) "the set includes the point $(0, 0)$". Which, in your opinion, will ultimately be topological properties.

6. Consider the "disk with a hole in the center", i.e., the set in the plane given by $x^2 + y^2 > 1$ and $x^2 + y^2 < 1$ What do you think its boundary and interior are? Do you think it is connected? bounded?



Problem Set 1 – Solutions



1. This set consists of two hyperbolas, as shown in tl figure.

2. i) Neither x^2 nor y^2 can be negative. So, we could have $x^2 + y^2 \leq 0$ only if $x^2 + y^2 = 0$, i.e., only if both x^2 and y^2 are zero, i.e., only if both x and y are zero. So, this is the set consisting only of the point $(0, 0)$. ii) In order that (x, y) not be in the set, all three of $x \geq 0, y \geq 0$, and $x + y \leq 0$ must fail. That is, x must be negative, y must be negative, and $x + y$ must be positive. But this is impossible. So, this set is the entire plane.

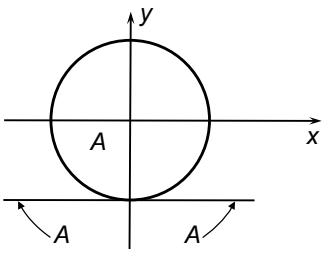
3. i) This does not describe a set. We are to be given (x, y), so what does "for some y" mean? ii) This does describe a set, and in fact it is the entire plane. iii) This does not describe a set. What is this z?

4. Your guess is as good as mine.

5. i) This problem will be a topological property, since it would seem to be unchanged under stretching and pulling. [In fact, it is, but it is a very complicated one to state.] ii) This probably will not be a topological property, since stretching and pulling could move a set away from the origin. [In fact, it is not.]

6. I would have guessed that its boundary will be the two circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$; its interior will be the set itself; and that it will be both connected and bounded. [In fact, these are precisely the conclusions one would draw by applying our definitions.]

Problem Set 2 Due: January 26, 1978



1. Let A be a "disk sitting on a line", i.e., the set of all points (x, y) with either $y = -1$ or $x^2 + y^2 < 1$. Find bnd (A) and int (A). Is A bounded? connected?

2. Find an example of a set which is both bounded and connected; one neither bounded nor connected; one bounded but not connected; one connected but not bounded.

3. Is the interior of every connected set connected? The boundary of every bounded set bounded?

4. Listed below are some proposed modifications of our definitions. In each case, apply the modified definition to the disk (i.e., for "boundary" or "interior" find, with the new definition, the boundary or interior of the disk; for "bounded" or "connected" determine, with the new definition whether the disk is bounded or connected). i) In the definition of "boundary", omit the word "positive". ii) In the definition of "boundary", change "... and also a point ..." to "... or a point ...". iii) In the definition of "interior", change "... for some positive number ..." to "... for any positive number ...". iv) In the definition of of "interior", omit the word 'positive". v) In the definition of "bounded", change "... some point p of the plane and some positive number c" to "every point p of the plane there exists a positive number c such that ...". vi) In the definition of "disconnected", omit "... and no point of A is in bnd (B)."

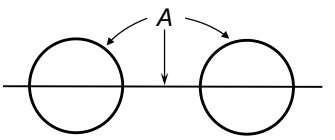
5. Find all sets A such that int (A) is the entire plane.

Problem Set 2 – Solutions

1. Bnd (A) consists of all points (x, y) with $x^2 + y^2 = 1$ or $y = -1$. Int (A) consists of all points (x, y) with $x^2 + y^2 < 1$. This set is not bounded, but it is connected.

2. A disk is both bounded and connected; two parallel lines are neither bounded nor connected; a set consisting of two points is bounded but not connected; the entire plane is connected but not bounded.

3. It is not true that the interior of every connected set is connected. Let, for example, A consist of two disks together with a line as shown. Then A is connected, but its interior is not. the boundary of every bounded set is bounded.



4. i) The boundary of the disk would be the empty set (since, for *every* ϵ one would have to find point q such that $d(p, q) \leq \epsilon$. But what about $\epsilon = -17$?) ii) The boundary of the disk would be the entire plane (since there will always, for any p, be either a q satisfying the first or a q satisfying the second). iii) The interior of the disk would be the empty set (for this is to hold for *every* positive ϵ. What about $\epsilon = 137$?) iv) The interior of the disk would be the entire plane (for the condition is satisfied for *some* ϵ, namely $\epsilon = -13$.) v) The disk would still be bounded. vi) The disk would be disconnected (choosing for C, say, all (x, y) with $x > 0$).

5. Every point of int (A) is a point of A. So, if int (A) is the entire plane, A must be the entire plane.

Problem Set 3 Due: January 30, 1978

1. Find all sets A in the plane such that bnd (A) is the empty set.

2. find all sets A in the plane every subset of which is connected.

3. Find a set A such that int (A) is the disk (all (x, y) with $x^2 + y^2 < 1$) and bnd (A) is the complement of this disk.

4. Is every set in the plane a subset of some connected set?

5. If bnd (A) is connected, need A be connected?

6. Can two different sets in the plane have the same boundary? The same interior?

7. Theorem. Let A be a bounded set in the plane. Then A^C is not bounded. [Give a formal proof.]

8. Theorem Let A and B be sets in the plane, with int $(A) = A$ and int $(B) = B$ Then int $(A \cap B) = A \cap B$. [Give a formal proof.]

9. Find disconnected sets A and B in the plane such that both $A \cup B$ and $A \cap B$ are connected.

10. Think of a subset A of the plane as being "dense" if every point of the plane is "intermingled" with the points of A; if no point of the plane is "out of direct contact" with the set A. [For example, the set of all (x, y) with x rational is to be "dense".] Find a suitable definition of "dense", and state and prove a few theorems.

Problem Set 3 – Solutions

1. The entire plane and the empty set have empty boundaries. Suppose that A were some other set with empty boundary. Then some point of the plane would be in A (since A is not empty), some point of the plane would be not in A (since A is not the entire plane), and no point of the plane would be in bnd (A) (since bnd (A) is empty). That is, we would have that the plane is disconnected. But the plane is connected. So, the only such A are the entire plane and the empty set.

2. If A contained two or more points, than one could form a subset of A consisting of just two points, and this subset would be disconnected. So, A must be either the empty set or a set with just one point.

3. Let A be the set of all points (x, y) in the plane such that either $x^2 + y^2 < 1$ or x is rational.

4. Yes. Every set in the plane is a subset of the entire plane, which is connected.

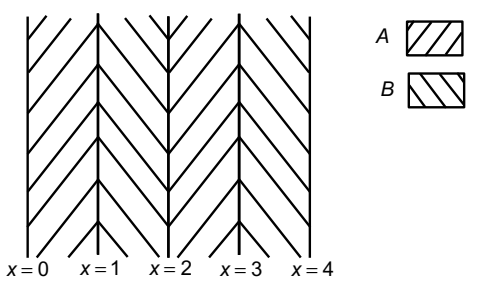
5. It need not. Our "two squares" set on page 15 has connected boundary, but is not connected.

6. Different sets, same boundary? Sure: For any set A in the plane, A and A^C are different, but have the same boundary, Different sets, same interior? Yes: Let A be a disk, and B a disk together with one additional point.

7. For contradiction, let A be bounded (so there is a point p and a positive number c such that every point of A is within c of p) and let A^C be bounded (so there is a point p' and a positive number c' such that every point of A^C is within c' of p'). Set $a = d(p, p')$. Then every point of A is within c of p, and so within $c + c' + a$ of p. every point of A^C is within c' of p', and so within $c' + a$ of p. and so within $c + c' + a$ of p. So, every point of the plane is within $c + c' + a$ of p: a contradiction.

8. [This is just a special case of the second theorem on page 21. It was a misprint: I meant the union,] i.e., int $(A \cup B) = A \cup B$. For this, a proof would be as follows.] That $A \cup B \subset \text{int}\,(A \cup B)$ follows from the theorem on page 23. That int $(A \cup B) \subset A \cup B$ follows from the theorem on page 18.

9. Let A consist of all (x, y) with $0 \leq x \leq 1$ or $2 \leq x \leq 3$, and let B consist of all (x, y) with $1 < x < 2$ or $3 < x < 4$. Then each is disconnected, while $A \cap B$ is the empty set (so, connected), and $A \cup B$ is a "wide vertical strip" (so, connected).



10. A set A in the plane is said to be *dense* if, for every point p of the plane and every positive number ϵ, there is a point q of A with $d(p, q) \leq \epsilon$. Typical theorems include: i) Let A and B be sets in the plane, with $A \subset B$ and A dense. Then B is dense. ii) The union of two dense sets in the plane is dense. iii) Let A be a dense set in the plane. Then int $(A) \cup$ bnd (A) is the entire plane.

Problem Set 4 Due: February 6, 1978

1. The curve given by $\gamma(t) = (t, t^2)$ is continuous at $t = 0$. For each of the following $\epsilon-$values, find a positive number δ such that, whenever $|t - 0| \leq \delta$, $d(\gamma(t), \gamma(0) \leq \epsilon$, $\epsilon = 3$; $1/2$; $1/100$.

2. The curve given by $\gamma(t) = (0, 0)$ if t is rational, $\gamma(t) = (1, 0)$ if t is irrational is not continuous at $t = 1$. Set $\epsilon = 1/2$. Show that there is no positive number δ such that, whenever $t - 1| \leq \delta$, $d(\gamma(t), \gamma(1) \leq \epsilon$.

3. Let γ be a continuous curve in the plane. Define a new curve, $\hat{\gamma}$, as follows. For t a number, let (x, y) be the point $\gamma(t)$, and then set $\hat{\gamma} = (x + 1, y)$. That is, $\hat{\gamma}$ is just γ "displaced one unit to the right along the $x-$axis". Show that $\hat{\gamma}$ is continuous. [Hint: Fix t_0. Since γ is continuous, for every positive ϵ there exists a positive δ such that ... for γ. Now let ϵ be a positive number. We must find positive $\hat{\delta}$ such that ... for $\hat{\gamma}$. What should we choose for this $\hat{\delta}$?]

4. Let X be the plane. For p and q points of X, set $d(p, q)$ = geometrical distance from p to $q + 1$. Is X, d a metric space? Why?

5. Let X be the plane. For p and q points of X, let $d(p, q)$ be the area of the triangle with vertices p, q and the origin, Is X, d a metric space? Why?

6. Let X be the set of positive integers. For p and q points of X (i.e., positive integers), set $d(p, q) = |p/q - 1|$. Is X, d a metric space? Why?

7. Let X, d and X, d' be metric spaces (i.e., two distances for the same set X). For p and q any two points of X, set $d''(p, q) = d(p, q) + d'(p, q)$. Is X, d'' a metric space? Why?

8. Let X be the plane. For p and q points of X, let $d(p, q)$ be the length of a side of the smallest square, with vertical and horizontal sides, which contains both p and q. Is X, d a metric space? Why?

Problem Set 4 – Solutions

1. The following are approximately the largest δ's which do the job (i.e., any smaller δ's are ok): 1.6, .45, .0009.

2. Since "1" is rational, $\gamma(1) = (0, 0)$. Let some positive ϵ be given. Then there is some irrational number t such that $|t - 1| \leq \delta$. But, for this t, $\gamma(t) = (1, 0)$, and so $d(\gamma(t), \gamma(0)) = 1$, which is not less than or equal to $1/2$. Thus, for every positive δ, we have found a t with $|t - 1| \leq \delta$, but such that $d(\gamma(t), \gamma(0)) \leq \epsilon$ fails.

3. Fix t_0, and let positive ϵ be given. Then, since γ is continuous at $t = t_0$, there is a positive δ such that, whenever $|t - t_0| \leq \delta$, $d(\gamma(t), \gamma(t_0)) \leq \epsilon$. But, by definition of $\hat{\gamma}$, we have $d(\hat{\gamma}(t), \hat{\gamma}(t_0)) = d(\gamma(t), \gamma(t_0))$. So, whenever $|t - t_0| \leq \delta$, $d(\hat{\gamma}(t), \hat{\gamma}(t_0)) \leq \epsilon$. We have thus shown that $\hat{\gamma}$ is continuous at $(t = t_0)$. Since t_0 is arbitrary, we have shown that $\hat{\gamma}$ is continuous.

4. No, $d(p, p) = 1$, not zero, so the first condition fails.

5. No. Let p, q, and the origin be arranged as shown. Then the triangle reduces to a line, and $d(p, q) = 0$, while $p \neq q$. The first condition fails.

6. No. It is not true that $d(p, q) = d(q, p)$ (e.g., for $p = 1$ and $q = 2$, so $d(p, q) = 1/2$ while $d(q, p) = 1$). Second condition fails.

7. Yes. each of the three conditions for d'' follows immediately from the corresponding conditions for d and d'.

8. Yes. the first two conditions are immediate. To check the third condition, one just has to play around with squares for a while to see that it is true.

Problem Set 5 Due: February 13, 1978

1. let X, d be a metric space. For p and q any two points of X, set $d'(p,q) = d(p,q)$ if $d(p,q) \leq 1$, and $d'(p,q) = 1$ if $d(p,q) > 1$. Show that X, d' is a metric space.

2. Let X, d be a metric space. Let X' be a set consisting of the points of X together with one additional point, x. For p and q in X, set $d'(p,q) = d(p,q)$; for p in X, set $d'(p,x) = 1$; set $d'(x,x) = 0$. Is X', d' a metric space? Why or why not?

3. Let γ be a continuous curve in the plane. Consider a new curve in the plane, $\hat{\gamma}$, given by: for t any real number, $\hat{\gamma}(t) = \gamma(2t)$. Thus, $\hat{\gamma}$ represents "the same path as γ, but traversed twice as fast". Show that $\hat{\gamma}$ is continuous.

4. Consider the following candidate for a set having more elements (as we have defined it) than the set of integers, but fewer elements than the set of numbers between 0 and 1: the set of numbers between 0 and $1/2$. Why does this set not work?

5. Let S, S', and S'' be sets, and suppose that S has the same number of elements as S', and S' the same number of elements as S'' (as we have defined "same number of elements"). Show that S has the same number of elements as S''.

Problem Set 5 – Solutions

1. Since the first two conditions are immediate, we have only to worry about the triangle inequality, $d'(p,q) + d'(q,r) \geq d'(p,r)$. Since in this metric space no distances exceed one, this could fail only if the left side were less than one (since certainly the right side cannot greater than one). But if the left side is less than one, we have $d'(p,q) = d(p,q)$ and $d'(q,r) = d(q,r)$ (since each is less than one), and so, by the triangle inequality in X, d, that $d(p,r)$ is also less than one. Hence, we must also have $d'(p,r) = d(p,r)$. Our triangle inequality now follows from that inequality in X, d.

2. This is not in general a metric space. Let p and r be points of X with $d(p,r) = 5$, say. Then the triangle inequality would demand $d(p,x) + d(x,r) \geq d(p,r)$, which fails.

3. Fix number t_0, and let positive ϵ be given. Then, since γ is continuous at $2t_0$, there is a positive δ such that, whenever $|t - 2t_0| \leq \delta$, $d(\gamma(t), \gamma(2t_0)) \leq \epsilon$. That is, setting $t' = 1/2$, whenever $|t' - t_0| \leq \delta/2$, $d(\gamma(2t'), \gamma(2t_0)) \leq \epsilon$. But, since $\hat{\gamma}(t) = \gamma(2t)$, this just means that $\hat{\gamma}$ is continuous at t_0).

4. But the set of numbers between 0 and $1/2$ has the same number of elements as the set of numbers between 0 and 1. Indeed, a correspondence is the following: With any number x between 0 and $1/2$, associate the number $2x$ (between 0 and 1). Every number of one set is made to correspond with some number in the other, and vice versa, and no numbers of either set are left over.

5. We have a correspondence between the elements of S and S' (such that every element of S corresponds to one of S', and vice versa, and such that no elements of S or S' are left over); and also a correspondence between the elements of S' and S'' (etc.). Now consider the following correspondence between the elements of S and S'': Given an element of X, find the element of S' to which it corresponds (under the first correspondence), and then the element of S'' to which that element of S' corresponds (under the second correspondence). One checks that this correspondence between S and S'' has all the required properties.

Problem Set 6 Due: February 20, 1978

1. Let X be the plane, d the usual geometrical distance, and d' twice that geometrical distance. Show that, for any set $A \subset X$, bnd (A) and int (A) are the same using d as d'.

2. Is it true that for A any subset of any metric space, with bnd $(A = A$, then int (A) is the empty set? Why or why not?

3. Let X be the "circle" metric space of page 34. Find all bounded subsets of X.

4. Let X, d be a metric space, with A having only a finite number of points. Prove that, for any $A \subset X$, int $(A) = A$ and bnd (A) is the empty set.

5. Is it true that, for any metric space X, d and for any subset A of X consisting of just two points, A is disconnected? Why or why not?

6. Let X, d and X, d' be metric spaces such that, for any $A \subset X$, int (A) (with respect to d) = int (A) (with respect to d'). Show that bnd (A) (respect to d) = bnd (A) (with respect to d').

Problem set 6 – Solutions

1. Let int and int' denote interiors using d and d', respectively. For p in int (A) there is a positive ϵ such that every point within $d-$distance of p is in A. But then every point within $d'-$distance 2ϵ would be in A, and so p would be in int (A). Similarly (using $1/2\epsilon$ instead of 2ϵ), every point of int' (A) is in int (A). That one also gets the same boundaries is immediate from the fifth theorem on page 48.

2. int (A) must be a subset of A, and can have no points in common with bnd (A). So, if bnd $(A) = A$, int (A) must be the empty set.

3. Every subset of X is bounded, because X itself is bounded. Indeed, given positive ϵ, let $p_1, \ldots, p_n$ be a finite number of points spaced less than ϵ apart around the circle. then every point of X is of course within ϵ of at least one of these points.

4. Let p be in A, and consider the numbers $d(p, q)$ for $q \neq p$. This is a finite number of positive numbers: Let ϵ be smaller than the smallest. Then every point within ϵ of p is in A (for we choose ϵ such that the only such point is p itself). So, p is in int (A). So, int $(A) = A$. Similarly, int $(A^C) = A^C$. Since the points of bnd (A) are precisely those points of X in neither int (A) nor int (A^C), bnd (A) must be the empty set (for every point is in either int $(A) = A$ or int $(A^C) = A^C$).

5. This is true. Let A consist of p and q, with $p \neq q$. Fix a positive number b less than $d(p, q)$, and let B be the subset of X consisting of all points r with $d(p, r) \leq b$. Then, of course, p is in B and q is not. We claim further that no point of A is in bnd (B), i.e., that neither p nor q is in bnd (B). Certainly p is not in bnd (B), for it is not true that there is a point within b of p and not in B, We further claim that q is not in bnd (B). Indeed, choose a positive number ϵ much that $b+\epsilon \leq d(p, q)$. (possible, since b is less than $d(p, q)$). Then no point within ϵ of q could be in B (i.e., within b of p), for were there such a point, r, the triangle inequality, $d(q, r) + d(r, p) \geq d(p, q)$, would contradict our choice of ϵ above. So, q is not in bnd (B). So, A is disconnected.

6. Since, by the fifth theorem on page 48, bnd $(A) = [\text{int}(A)]^C \cap [\text{int}(A^C)]^C$, it is immediate that two metrics having the same interiors also have the same boundaries. [Also, is fact, same connected set.]

Problem Set 7 Due: February 27, 1978

1. i) Let X, d be a metric space, and p and q points of X with $p \neq q$. Show that there is a self-interior set A with p in A and q not in A.

ii) Let X, d be a metric space, and $X, \mathcal{J}$, its underlying topological space, and p and q points of X with $p \neq q$. Show that there is an open set A in this topological space with p in A and q not in A.

iii) Let X be the plane with the indiscrete topology, and p and q points of X with $p \neq q$. Show that there is an open set A in this topological space with p in A and q not in A.

iv) Show that the plane with the indiscrete topology is not the underlying topological space of any metric space.

2. Let X be a set with just three points. Find all $\mathcal{J}$'s (i.e., all collections of subset of X) such that $X, \mathcal{J}$ is a topological space. [Hint: There are exactly twenty-nine of them.]

3. Let X, d be a metric space, with X finite. Show that the the underlying topological space is discrete.

4. Let X be a set, and $\mathcal{J}$ a collection of subsets of X such that the second and third conditions for a topological space are satisfied. Is it true that, if there is now included in the collection $\mathcal{J}$ also X itself and the empty set, then the result will always be a topological space? Why or why not?

Problem set 7 – Solutions

1. i) Let $d(p, q) = a$, a positive number. Let A be the set of all points r with $d(p, r) < a/2$. Then clearly p is in A and q is not in A. By the theorem on page 58, A is self-interior. ii) Since the open sets in the underlying topological space are the self-interior sets, this follows from i). iii) In the indiscrete topology, the only open sets are X and the empty set. But A empty will not do, for p will not be in A; and $A = X$ will not do, for q will be in A. iv) Let X, d be any metric space with X the plane, and let $p \neq q$. By ii), these is an open set, in the underlying topology, containing p but not q: by iii) there is no open set, in the indiscrete topology, containing p but not q. So these topologies must be different.

2. One just has to hunt them down. There is the indiscrete (1) and the discrete (1). The rest are: $[x, y, xy]$ (3) [Which means: the open sets are the empty, X itself, the set consisting of just x, the set consisting of just y, and the set consisting of just x and y; and the total number of such possibilities, when rearrangements of x, y, and z are included, is 3], $[xy]$(3), $[x]$(3), $[x, yz]$(3), $[x, xy]$(6), $[x, xy, xx]$(3), $[x, y, xy, xz]$(6).

3. We have seen that, in a metric space with X finite, int $(A) = A$ for every $A \subset X$, i.e., that every A is self-interior. So, in the underlying topological space, every subset of X is open. That is, the underlying topological space is discrete.

4. This is true. Let $\mathcal{J}'$ denote $\mathcal{J}$, together with $\emptyset$ (empty set), and X itself. 1. X and $\emptyset$ are certainly in $\mathcal{J}'$. 2. Let A and B be in $\mathcal{J}'$. If either is empty, then $A \cap B = \emptyset$, and is in $\mathcal{J}'$. If either is X, then $A \cap B$ is the other, and so is in $\mathcal{J}'$. If neither is empty or X, then each of A and B is in $\mathcal{J}$ and so $A \cap B$ is in $\mathcal{J}$, and so $A \cap B$ is in $\mathcal{J}'$. 3. Let C be the union of any collection of sets in the collection $\mathcal{J}'$. We might as well leave out empty sets in this union, since they will not change C. If any of these sets is X, then $C = X$, and so is in $\mathcal{J}'$. If none is X, then they are in $\mathcal{J}$, and so the union is in $\mathcal{J}$, and so that union, C is in $\mathcal{J}'$.

Problem Set 8 Due: March 8, 1978

1. Let X, $\mathcal{J}$ be a topological space in which every subset of X consisting of a single point is open. Prove that this topological space is discrete.

2. Consider the topological space of pages 56 – 56. Find the connected subsets of X.

3. Let X be the plane, and let the open sets be the empty set and all subset of X containing the origin. Let A consist of all (x, y) with $x^2 + y^2 < 1$.

i) prove that this is a topological space.

ii) Find int (A).

iii) Find bnd (A).

iv) Is A a connected? Why or why not?

4. Let X, $\mathcal{J}$ be a topological space, and A a connected subset of X. Prove that $A \cup$ bnd (A) is connected. [Note that one would have expected this intuitively!] [Hints: Suppose disconnected, so there is a set C such that some point, p, of $A \cup$ bnd (A) is in C, some point, q, of $A \cup$ bnd (A) is not in C, and no point of $A \cup$ bnd (A) is in bnd (C). Now it could hardly be that both p and q are in A (as opposed to bnd (A)), for that would mean that A would be disconnected. Suppose, then, that p is in bnd (A). Now use i) the fact that p is in C, ii) the fact that p is in bnd (C) (since no point of $A \cup$ bnd (A) is in bnd (C)), and iii) the fact that p is in bnd (A) to show that there must be a point p' in A and C. Similarly, if q were in bnd (A), show that there must be a point q' in A and not in C. Thus, we wind up finally with a point (p or p') in A and in C, a point (q, or q') in A and not in C, and with no point in A in bnd (C). But this would imply that A is disconnected.]

Problem set 8 – Solutions

1. In a topological space, any union of open sets is open. But any subset of X can be written as a union of sets consisting of a single point. So, if these are open, the space is discrete.

2. For any subset B of X, either int (B) is empty, or int (B^C) is empty. So how, given A, could one find B such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B) (since if no point of A is in bnd (B), the point of A in B must be in int (B), and the point of A not in B must be in int (B^C))? Every subset of X is connected.

3. i) 1. The empty set and X itself are open. 2. The intersection of two sets containing the origin contains the origin. 3. The union of any collection of sets containing the origin contains the origin. ii) int $(A) = A$. Indeed, this A is open, since it contains the origin. iii) bnd $(A) = A^C$. Indeed, let p be in A^C. Then every open neighborhood of p contains a point not in A (namely, p), and also a point in A (namely, the origin). iv) In order that A be disconnected, we must find B such that some point of A is in B, some point of A is not in B, and no point of A is in bnd (B). But, if B contains the origin, then bnd $(B) = B^C$ – so, the point of A not in B is in bnd (B); if B does not contain the origin, then bnd $(B) = B$ – so, the point of A in B is in bnd (B). In either case, there would be a point of A in bnd (B). So, we shall never find set B which shows disconnected. This A is connected.

4. See notes, pages, 67 and 67.

Special Topics in Particle Physics

Robert Geroch

April 18, 2005

Contents

1 The Klein-Gordon Equation **3**

2 Hilbert Space and Operators **6**

3 Positive-Frequency Solutions of the Klein-Gordon Equation **8**

4 Constructing Hilbert Spaces and Operators **13**

5 Hilbert Space and Operators for the Klein-Gordon Equation **15**

6 The Direct Sum of Hilbert Spaces **18**

7 The Completion of an Inner-Product Space **20**

8 The Complex-Conjugate Space of a Hilbert Space **22**

9 The Tensor Product of Hilbert Spaces **23**

10 Fock Space: The Symmetric Case **28**

11 Fock Space: The Anti-Symmetric Case **31**

12 Klein-Gordon Fields as Operators **34**

13 The Hilbert Space of Solutions of Maxwell's Equations **39**

14 Maxwell Fields as Operators **44**

15 The Poincaré Group **47**

16 Representations of the Poincaré Group **49**

17 Casimir Operators: Spin and Mass **54**

18 Spinors **61**

19 The Dirac Equation **65**

20 The Neutrino Equation **72**

21 Complex Klein-Gordon Fields **75**

22 Positive Energy **75**

23 Fields as Operators: Propagators **79**

24 Spin and Statistics **83**

25 $\star$-Algebras **85**

26 Scattering: The S-Matrix **91**

27 The Hilbert Space of Interacting States **94**

28 Calculating the S-Matrix: An Example **96**

29 The Formula for the S-Matrix **100**

30 Dimensions **102**

31 Charge Reversal **104**

32 Parity and Time Reversal **107**

33 Extending Operators to Tensor Products and Direct Sums **111**

34 Electromagnetic Interactions **114**

35 Transition Amplitudes **119**

1 The Klein-Gordon Equation

We want to write down some kind of a quantum theory for a free relativistic particle. We are familiar with the old Schrödinger prescription, which more or less instructs us as to how to write down a quantum theory for a simple, nonrelativistic classical system. The idea is to mimic as much at that prescription as we can. In doing this, a number of difficulties will be encountered which, however, we shell be able to resolve. There is a reasonable and consistent quantum theory for a free relativistic (spin zero) particle.

Recall the Schrödinger prescription. We have a classical system (e.g., a pendulum, or a ball rolling on a table). The manifold of possible instantaneous configurations of this system is called *configuration space*, and points of this manifold are labeled by letters such as x. However, in order to specify completely the state of the system (i.e., in order to give enough information to uniquely determine its future evolution), we must specify at some initial time both its configuration x and its momentum p. The collection of such pairs (x, p) is called *phase space.* (More precisely, phase space is the cotangent bundle of configuration space.) Finally, the dynamics of the system is described by a certain real-valued function on phase space, $H(x, p)$, the *Hamiltonian.* The time-evolution of the system (i.e., its point in phase space) is given by Hamilton's equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}x = \frac{\partial}{\partial p}H \qquad \frac{\mathrm{d}}{\mathrm{d}t}p = -\frac{\partial}{\partial x}H \tag{1}$$

Thus, the complete dynamical history of the classical system is represented by curves (solutions of Eqn. (1)), $(x, p)(t)$, in phase space. (More precisely, by integral curves of the Hamiltonian vector field in phase space.)

The state of the corresponding quantum system is characterized not by a point in phase space as in the classical case, but rather by a complex-valued function $\psi(x)$ on configuration space. The time-evolution of the state of the system is then given, not by Eqn. (1) as in the classical case, but rather by the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\psi = H\left(x, -i\hbar\frac{\partial}{\partial x}\right)\psi \tag{2}$$

where the operator on the right means "at each appearance of p in H, substitute $-i\hbar\frac{\partial}{\partial x}$". (Clearly, this "prescription" may become ambiguous for a sufficiently complicated classical system.) Thus, the complete dynamical history of the system is represented by a certain complex-valued function $\psi(x, t)$ of location in configuration space and time.

We now attempt to apply this prescription to a free relativistic article of mass $m \geq 0$. The (4-)momentum of such a classical particle, p^a, satisfies $p^a p_a = m^2$. (Latin indices represent (4-)vectors or tensors in Minkowski space. We use signature $(+, -, -, -)$.) Choose a particular unit (future-directed) timelike vector t^a (a "rest frame"), and consider the component of p^a parallel to t^a, $E = p^a t_a$, and its components perpendicular to t^a, $\vec{p}$. Then, from $p^a p_a = m^2$,

we obtain the standard relation between this "energy" and "3-momentum":

$$E = \left(\vec{p}\cdot\vec{p} + m^2\right)^{1/2}. \tag{3}$$

(Here and hereafter, we set the speed of light, c, equal to one.) The plus sign on the right in Eqn. (3) results from the fact that p^a is a future-directed timelike vector. It seems natural to consider Eqn. (3) as representing the "Hamiltonian" for a free relativistic particle. We are thus led to consider the dynamical history of the quantum particle as being characterized by a complex-valued function $\phi(x^a)$ on Minkowski space (x^a represents position in Minkowski space — it replaces both the "x" and "t" in the Schrödinger theory), satisfying the equation:

$$i\hbar\frac{\partial}{\partial t}\phi = \left[-\hbar^2\nabla^2 + m^2\right]^{1/2}\phi \tag{4}$$

The first set of difficulties now appear. In the first place, it is not obvious that Eqn. (4) is in any sense Lorentz invariant - that it is independent of our original choice of p^a. Furthermore, it is not clear what meaning is to be given to the operator on the right side of Eqn. (4): what does the "square root" of a differential operator mean? Both of these difficulties can be made to disappear, after a fashion, by multiplying both sides of Eqn. (4) by another, equally obscure, operator, $i\hbar\frac{\partial}{\partial t} + \left[-\hbar^2\nabla^2 + m^2\right]^{1/2}$, and expanding using associativity. The result is the *Klein-Gordon* equation:

$$\left(\Box + \frac{m^2}{\hbar^2}\right)\phi = 0, \qquad \text{or} \qquad \left(\Box + \mu^2\right)\phi = 0 \tag{5}$$

which is both meaningful and relativistically invariant. (We set $\mu = m/\hbar$.) We might expect intuitively that the consequence of multiplying Eqn. (4) by something to get Eqn. (5) will be that the number of solutions of Eqn. (5) will be rather larger than the number of solutions of Eqn. (4) (whatever that means.) As we shall see later, this intuitive feeling is indeed borne out.

To summarize, we have decided to describe our quantized free relativistic particle by a complex-valued function ϕ on Minkowski space, which satisfies Eqn. (5).

Just for the fun of it, let's look for a solution of Eqn. (5). We try

$$\phi = e^{ik_a x^a} \tag{6}$$

where k_a is a constant vector field in Minkowski space. Substituting Eqn. (6) into Eqn. (5), we discover that (6) is indeed a solution provided

$$k^a k_a = \mu^2 \tag{7}$$

i.e., provided k_a is timelike with norm μ.

In the Schrödinger prescription, the wave function ψ has a definite and simple physical interpretation: $\psi\psi^*$ represents the probability contribution for finding the particle. What is the analogous situation with regard to the solutions of the

Klein-Gordon equation? We know, e.g., from electrodynamics, that what is a "density" in a nonrelativistic theory normally becomes "the time-component of a 4-vector" in a relativistic theory. Thus, to replace "$\psi\psi^*$", we are led to look for some 4-vector constructed out of a solution of the Klein-Gordon equation. This suggestion is further strengthened by the observation that for a Schrödinger particle in a potential (so $H = (1/2m)p^2 + V(x)$), we have the equation

$$\frac{\partial}{\partial t}\left(\psi\psi^*\right) = -\vec{\nabla}\cdot\left[\frac{\hbar}{2mi}\left(\psi^*\vec{\nabla}\psi - \psi\vec{\nabla}\psi^*\right)\right] \tag{8}$$

(Proof: evaluate the time-derivatives on the left using (2), and verify that the result is the same as the expression on the right.) This looks very much like the nonrelativistic form of the statement that the 4-divergence of some 4-vector vanishes. Hence, we want to construct some divergence-free 4-vector from solutions of the Klein-Gordon equation. One soon discovers such an object which, in fact, looks suggestively like the object appearing in Eqn. (8):

$$J_a = \frac{1}{2i}\left(\phi^*\nabla_a\phi - \phi\nabla_a\phi^*\right) \tag{9}$$

Note that, because of (5), J_a is divergence-free.

We cannot interpret the "time-component" of Eqn. (9) as a probability density for the particle unless this quantity is always nonnegative, that is to say, that $J_a t^a \geq 0$ for every future-directed timelike vector t^a, that is to say, unless J_a itself is future-directed and timelike. To see whether this is indeed the case, we evaluate J_a for the plane-wave solution (Eqn. (6)), and find:

$$J_a = k_a \tag{10}$$

This expression is indeed timelike, but is not necessarily future-directed: Eqn. (6) is a solution of the Klein-Gordon equation whether k_a is future- or past-directed. Thus, we have not succeeded in interpreting a solution of the Klein-Gordon equation in terms of a "probability density for finding the particle."

We next consider the situation with regard to the initial value problem. Since the Schrödinger equation is first order in time derivatives, a solution of that equation is uniquely specified by giving $\psi'(x)$ at some initial time, say $t = 0$. The Klein-Gordon equation, on the other hand, is second order in time derivatives. Hence, to specify a solution, one must give both ϕ and $\partial\phi/\partial t$ at the initial time $t = 0$. This radical change in the structure of the initial data is clearly a consequence of our having "squared" Eqn. (4). It is still another indication that the transition to the relativistic case is not just a trivial application of the Schrödinger prescription.

Finally, let's look briefly at the structure of the space of solutions of the Klein-Gordon equation. In the non-relativistic case, the space of solutions of the Schrödinger equation is a Hilbert space: it's obviously a complex vector space, and we define the norm of the state ψ by:

$$\|\psi\|^2 = \int_{t=\text{const.}} \psi\psi^* \mathrm{d}V \tag{11}$$

That the real number (11) is independent of the t = const. surface over which the integral is performed is a consequence of Eqn. (8) (assuming, as one always does, that ψ falls off sufficiently quickly at infinity.) One might therefore be tempted to try to define the norm of a solution of the Klein-Gordon equation as an integral of J_a

$$\int_S J_a \mathrm{d}s^a \tag{12}$$

over a spacelike 3-plane S. But it is clear from (10) that the expression (12) will not in general be nonnegative. Thus, the most obvious way to make a Hilbert space out of solutions of the Klein-Gordon equation fails. This, of course, is rather embarrassing, for we are used to doing quantum theory in a Hilbert space, with Hermitian operators representing observables, etc.

To summarize, a simple "relativization" of the Schrödinger equations leads to a number of maladies.

2 Hilbert Space and Operators

The collection of states of a quantum system, together with certain of the structure naturally induced on this collection, is described by a mathematical object known as a Hilbert space. We recall the basic definitions.

A *Hilbert space* consists, first of all, of a set H. Secondly, H has the structure of an Abelian group. That is to say, given any two elements, ξ and η, of H, there is associated a third element of H, written $\xi + \eta$, this operation subject to the following conditions:

H1. For $\xi, \eta \in H$, $\xi + \eta = \eta + \xi$.

H2. For $\xi, \eta, \phi \in H$, $(\xi + \eta) + \phi = \xi + (\eta + \phi)$.

H3. There is an element of H, written "0", with the following property: for each $\xi \in H$, $\xi + 0 = \xi$.

H4. If $\xi \in H$, there exists an element of H, written "$-\xi$", with the following property: $\xi + (-\xi) = 0$.

Furthermore, H has the structure of a complex vector space. That is to say, with each complex number α and each element ξ of H there is associated an element of H, written $\alpha\xi$, this operation subject to the following conditions:

H5. For $\xi, \eta \in H$, $\alpha \in \mathbb{C}$, $\alpha(\xi + \eta) = \alpha\xi + \alpha\eta$.

H6. For $\xi \in H$, $\alpha, \beta \in \mathbb{C}$, $(\alpha + \beta)\xi = \alpha\xi + \beta\xi$ and $(\alpha\beta)\xi = \alpha(\beta\xi)$.

H7. For $\xi \in H$, $1\xi = \xi$.

There is, in addition, a positive-definite inner product defined on H. That is to say, with any two elements, ξ and η, of H there is associated a complex number, written (ξ, η), this operation subject to the following conditions:

H8. For $\xi, \eta, \phi \in H$, $\alpha \in \mathbb{C}$, $(\alpha\xi + \eta, \phi) = \alpha(\xi, \phi) + (\eta, \phi)$.

H9. For $\xi, \eta \in H$, $\overline{(\xi, \eta)} = (\eta, \xi)$.

H10. For $\xi \in H$, with $\xi \neq 0$, $(\xi, \xi) > 0$. (That (ξ, ξ) is real follows from **H9**.)

We sometimes write $\|\xi\|$ for $\sqrt{(\xi, \xi)}$. Finally, we require that this structure have a property called completeness. A sequence, ξ_i $(i \in 1, 2, ...)$, of elements of H is called a *Cauchy sequence* if, for every number $\epsilon > 0$, there is a number N such that $\|\xi_i - \xi_j\| < \epsilon$ whenever i and j are greater than N. A sequence is said to *converge* to $\xi \in H$ if $\|\xi - \xi_j\| \to 0$ as $i \to \infty$. H is said to be *complete* if every Cauchy sequence converges to an element of H.

H11. H is complete.

There are, of course, hundreds of elementary properties of Hilbert spaces which follow directly from these eleven axioms.

A (linear) operator on a Hilbert space H is a rule A which assigns to each element ξ of H another element of H, written $A\xi$, this operation subject to the following condition:

O1. $\xi, \eta \in H$, $\alpha \in \mathbb{C}$, $A(\alpha\xi + \eta) = \alpha A\xi + A\eta$.

We shall discuss the various properties and types of operators when they arise.

There is a fundamental difficulty which arises when one attempts to use this mathematical apparatus in physics. The "collection of quantum states" which arises naturally in a physical problem normally satisfies **H1**–**H10**. (This is usually easy to show in each case.) The problem is with **H11**. The most obvious collection of states often fails to satisfy the completeness condition. As one wants a Hilbert space, he normally corrects this deficiency by completing the space, that is, by including additional elements so that all Cauchy sequences will have something to converge to. (There is a well-defined mathematical procedure for constructing, from a space which satisfies **H1**–**H10**, a Hilbert space.) The unpleasant consequence of being forced to introduce these additional states is that the natural operators of the problem, which were defined on the original collection of states, cannot be defined in any reasonable way on the entire Hilbert space. Thus, they are not operators at all as we have defined them, for they only operate on a subset of the Hilbert space. Fortunately, this subset is dense. (A subset D of a Hilbert space H is said to be dense if, for every element ξ of H, there is a sequence consisting of elements of D which converges to ξ.) Some very unaesthetic mathematical techniques have been devised for dealing with such situations. (See Von Neumann's book on Mathematical Foundations of Quantum Mechanics.)

This problem is not confined to quantum field theory. It occurs already in Schrödinger theory. For example, the collection of smooth solutions of the Schrödinger equation for which the integral (11) converges satisfy **H1**–**H10**, but not **H11**. To complete this space, we have to introduce "solutions" which are, for example, discontinuous. How does one apply the Schrödinger momentum operator to such a wave function?

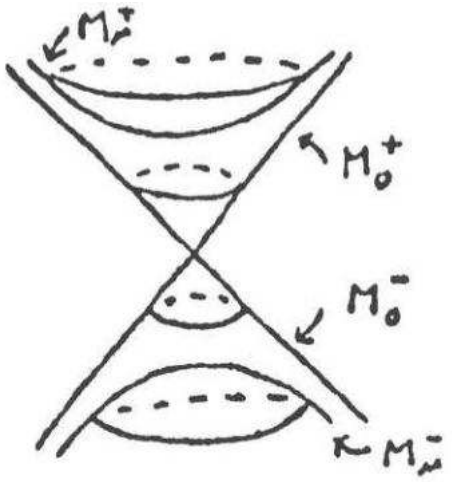


Figure 1: The mass shell in momentum space.

3 Positive-Frequency Solutions of the Klein-Gordon Equation

We represent solutions of the Klein-Gordon equation as linear combinations of plane-wave solutions (Eqn. (6)):

$$\phi(x) = \int_{M_\mu} f(k_a) e^{ik_a x^a} \mathrm{d}V_\mu \tag{13}$$

Of course, we wish to include in the integral (13) only plane-waves which satisfy the Klein-Gordon equation, i.e., only plane waves whose k_a satisfy the normalization condition (7). The four-dimensional (real) vector space of constant vector fields in Minkowski space-time is called *momentum space*. The collection of all vectors k_a in momentum space which satisfy Eqn. (7) consists of two hyperbolas (except in the case $\mu = 0$, in which case the hyperbolas degenerate to the two null cones through the origin). This collection is called the *mass shell* (associated with μ), M_μ (Fig. 1). Thus the function f in (13) is defined only on the mass shell, and the integral is to be carried out over the mass shell. It is convenient, furthermore, to distinguish the *future mass shell* M_μ^+ (consisting of future-directed vectors which satisfy (7)) and the *past mass shell* M_μ^- (consisting of past-directed vectors which satisfy (7)), so $M_\mu = M_\mu^+ \cup M_\mu^-$.

Eqn. (13) immediately suggests two questions: i) What are the necessary and sufficient conditions on the complex-valued function f on M_μ in order that the integral (13) exist for every x^a, and in order that the resulting $\phi(x)$ be smooth and satisfy the Klein-Gordon equation? ii) What are the necessary and sufficient conditions on a solution $\phi(x)$ of the Klein-Gordon equation in

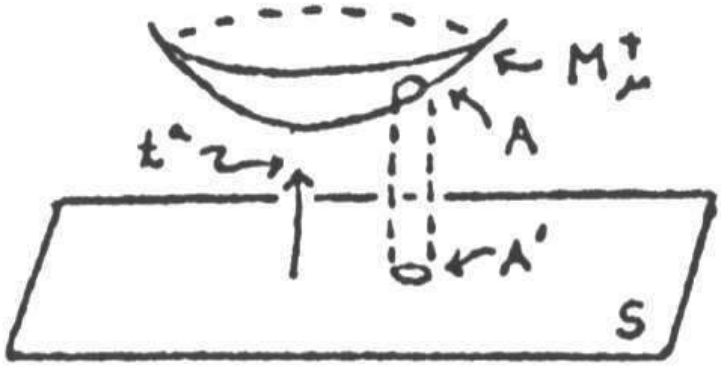


Figure 2: The volume element on the mass shell.

order that it can be expressed in the form (13) for some f? These, of course, are questions in the theory of Fourier analysis. It suffices for our purposes, however, to remark that the required conditions are of a very general character (that functions not be too discontinuous, and that, asymptotically, they go to zero sufficiently quickly). The point is that all the serious things we shall do with the Klein-Gordon equation will be in momentum space. We shall use Minkowski Space and $\phi(x)$ essentially only to motivate definitions and constructions on the f's in momentum space.

One question regarding (13) which must be answered is what is the volume element $\mathrm{d}V_\mu$ we are using on the mass shell. Of course, it doesn't make any real difference, for a change in the choice of volume element would merely result in a suitable readjustment of the f's. Our choice can therefore be dictated by convenience. We require that our volume element be invariant under Lorentz transformations on momentum space (note that these leave the mass shell invariant), and that it be applicable also in the case $\mu = 0$. It is easy to state an appropriate volume element in geometrical terms. Let $\mu > 0$. Then the mass shell is a spacelike 3-surface in momentum space, in which there is a metric, so a metric is induced on M_μ. A metric on this 3-manifold defines a volume element $\mathrm{d}\tilde{V}_\mu$. This $\mathrm{d}\tilde{V}_\mu$ is clearly Lorentz-invariant, but, unfortunately, it approaches zero as $\mu \to 0$. To correct this, we define

$$\mathrm{d}V_\mu = \mu^{-1}\mathrm{d}\tilde{V}_\mu \tag{14}$$

which is easily verified to he nonzero also on the null cone. In more conventional terms, our volume-element can be described as follows. Choose a unit time-like vector t^a in momentum space, and let S be the spacelike 3-plane perpendicular to t^a. Then any small patch A on M_μ, located at the point k_a, can be projected along t^a to give a corresponding patch A' on S (see Fig. 2.) Let $\mathrm{d}V'_\mu$ be the volume of A' on S (using the usual volume element in the 3-space S). Our

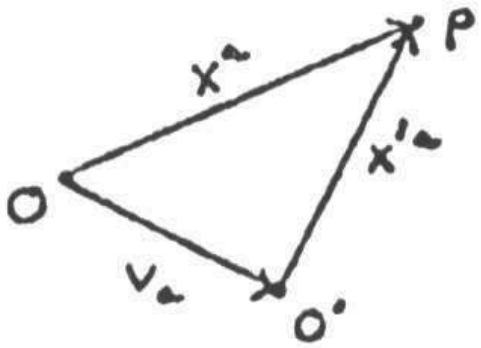


Figure 3: Illustration of vectors in Minkowski space.

volume element on M_μ is then given by the following expression:

$$\mathrm{d}V_\mu = \mathrm{d}V'_\mu |t^a k_a|^{-1} \tag{15}$$

The existence of a limit as $\mu \to 0$ is clear from (15), but Lorentz-invariance (independence of the choice of t^a) is not.

Is there any "gauge" in f? Given a solution $\phi(x)$ of the Klein-Gordon equation, is f uniquely determined by (13)? The only arbitrary choice which was made in writing (13) was the choice of an origin: "x^a" refers to the position vector of a point in Minkowski space with respect to a fixed origin. We are thus led to consider the behavior of the f's under origin changes. Let O and O' be two origins, and let v_a be the position vector of O' relative to O (see Fig. 3.) Let the position vectors of a point p in Minkowski space with respect to 0 and O' be x^a and x'^a, respectively, whence

$$x'^a = x^a - v^a \tag{16}$$

Then, if f and f' are the Fourier transforms of ϕ with respect to the origins O and O', respectively, we have

$$\phi(p) = \int\limits_{M_\mu} f e^{ik_a x^a} \mathrm{d}V_\mu = \int\limits_{M_\mu} f' e^{ik_a x'^a} \mathrm{d}V_\mu \tag{17}$$

Clearly, we must have

$$f'(k) = f(k) e^{ik_a v^a} \tag{18}$$

Thus, when we consider states as represented by functions on the mass shell, it is necessary to check that conclusions are unchanged if (18) is applied simultaneously to all such functions.

Now look again at Eqn. (3). It says, in particular, that the energy-momentum vector is future-directed. This same feature shows up in the right side of Eqn. (4) by the plus sign. If this sign were replaced by a minus, we would be dealing with a past-directed energy-momentum vector. The trick we used to obtain Eqn. (5) from (4) amounted to admitting also past-directed energy-momenta. It is clear now how Eqn. (4) itself can be carried over into a well-defined and fully relativistic condition on ϕ. We merely require that the f of Eqn. (13) vanish on M_μ^-. We call the corresponding solutions of the Klein-Gordon equation *positive-frequency* (or positive-energy) solutions. Defining negative-frequency solutions analogously, it is clear that every solution of the Klein-Gordon equation (more precisely, every solution which can be Fourier-analyzed) can be written uniquely as the sum of a positive-frequency and a negative-frequency solution. The positive-frequency solutions (resp., negative-frequency solutions) form a vector subspace of the vector space of all (Fourier analyzable) solutions.

To summarize, we are led to take as the "wave function of a free relativistic (spin zero) particle" a positive-frequency solution of the Klein-Gordon equation. To what extent does this additional condition take care of the difficulties we discovered in Sect. 1?

Consider first the current, Eqn. (9). We saw before, from the example of a negative-frequency plane wave, that in general J is neither future-directed or even timelike. Is it true that J is future-directed timelike for a positive-frequency solution of the Klein-Gordon equation? Unfortunately, the answer is no. Consider a linear combination of two positive-frequency plane waves:

$$\phi = e^{ik_a x^a} + \alpha e^{ik'_a x^a} \tag{19}$$

That is, α is a complex constant, and k_a and k'_a are future-directed constant vectors satisfying (7). (Strictly speaking, this example is not applicable, for (19) cannot be Fourier analyzed. It is not difficult, however, to appropriately smear (19) over the future mass shell to obtain an example without this deficiency.) Substituting (19) into (9), we obtain:

$$\begin{aligned} J_a = \frac{1}{2}k_a \left[2 + \alpha e^{i(k'_b - k_b)x^b} + \alpha^* e^{i(k_b - k'_b)x^b}\right] \\ + \frac{1}{2}k'_a \left[2\alpha\alpha^* + \alpha e^{i(k'_b - k_b)x^b} + \alpha^* e^{i(k_b - k'_b)x^b}\right] \end{aligned} \tag{20}$$

Clearly, one can choose α, k_a, and k'_a so that this J_a is not timelike in certain regions. Thus, even the assumption of positive-frequency solutions does not resolve the difficulty associated with not having a simple probabilistic interpretation for our wavefunction ϕ: we still cannot think of $J_a t^a$ (with t^a unit, future-directed, timelike) as representing a probability density for finding the particle. The resolution of this problem must await our introduction of a position operator.

Note from Eqn. (20) that J_a is trying very hard to be timelike and future-directed in the positive-frequency case: it is only the cross terms between the two plane waves which destroys this property. This observation suggests that, in

the positive-frequency case, the integral of J_a over a spacelike 3-plane might be positive. In order to check on this possibility, we want to rewrite the integral of J_a in terms of the corresponding function f on M_μ. It will be more illuminating to do this for the general solution ϕ of the Klein-Gordon equation, i.e., not assuming, for the time being, that ϕ is positive-frequency. Substituting (13) into (9):

$$J_a = \int\limits_{M_\mu} \mathrm{d}V_\mu \int\limits_{M_\mu} \mathrm{d}V'_\mu \, \frac{1}{2} k'_a \left[f^*(k) f(k') e^{i(k'_b - k_b)x^b} + f(k) f^*(k') e^{i(k_b - k'_b)x^b} \right] \quad (21)$$

We now let S be a spacelike 3-plane through the origin, and let t^a be the unit, future-directed normal to S. Then

$$\begin{aligned} \int\limits_S J_a t^a \, \mathrm{d}S = \int\limits_{M_\mu} \mathrm{d}V_\mu \int\limits_{M_\mu} \mathrm{d}V'_\mu \, \frac{1}{2} k'_a t^a \Bigg[& f^*(k) f(k') \int\limits_S e^{i(k'_b - k_b)x^b} \, \mathrm{d}S \\ & + f(k) f^*(k') \int\limits_S e^{i(k_b - k'_b)x^b} \mathrm{d}S \Bigg] \end{aligned} \quad (22)$$

But, from the theory of Fourier analysis

$$\int\limits_S e^{i(k'_b - k_b)x^b} \, \mathrm{d}S = (2\pi)^3 |t^a k_a|^{-1} \delta(k - k') \quad (23)$$

and so (22) becomes

$$\int\limits_S J_a t^a \, \mathrm{d}S = (2\pi)^3 \left[\int\limits_{M_\mu^+} f f^* \, \mathrm{d}V_\mu - \int\limits_{M_\mu^-} f f^* \, \mathrm{d}V_\mu \right] \quad (24)$$

In particular, if f vanishes on M_μ^-, the left side of (24) must be positive, or vanishes if and only if ϕ vanishes. (The actual form of (24) was rather suggested by Eqn. (15). If we had omitted the absolute-value sign on the right (and what better thing could be done with an absolute-value sign?), the "volume element" on M_μ^- would have been negative.) This calculation was not done merely for idle curiosity; the right side of (24) will be important later.

We saw before that the initial-value problem for the Klein-Gordon equation is as follows: one must specify ϕ and $t^a \nabla_a \phi$ on an initial spacelike 3-plane. How does the initial-value problem go for positive-frequency solutions of the Klein-Gordon equation? In fact, we only have to specify ϕ as initial data in this case. To see this, suppose we know the value of the integral

$$\phi(x) = \int\limits_{M_\mu^+} f(k) e^{ik_a x^a} \, \mathrm{d}V_\mu \quad (25)$$

for every x^a which is perpendicular to a unit timelike vector t^a at the origin (i.e., on the spacelike 3-plane perpendicular to t^a, through the origin). The integral (25) can certainly be expressed as a Fourier integral over S (t^a sets up a one-to-one correspondence between M_μ^+ and S). But then, by taking a Fourier transform, we can determine f on M_μ^+. Thus, we know ϕ throughout Minkowski space. That is to say, when we properly interpret (4), we obtain a "Schrödinger-type" initial-value problem. If we ignore questions of smoothness and convergence of Fourier integrals, the situation can be roughly summarized as follows:

1. There is a one-to-one correspondence between: i) solutions of the Klein-Gordon equation, ii) complex-valued functions f on M_μ, and iii) values of ϕ and $t^a \nabla_a \phi$ on a spacelike 3-plane.

2. There is a one-to-one correspondence between: i) positive-frequency solutions of the Klein-Gordon equation, ii) complex-valued functions on M_μ^+, and iii) values of ϕ on a spacelike 3-plane.

4 Constructing Hilbert Spaces and Operators

There is a general and extremely useful technique for obtaining a Hilbert space along with a collection of operators on it. It is essentially this technique which is used, for example, in treating the Schrödinger and Klein-Gordon equations. It is convenient, therefore, to describe this construction, once and for all, in a general case. Special cases can then be treated as they arise.

The fundamental object we need is some n-dimensional manifold M on which there is specified a smooth, nowhere-vanishing volume-element $\mathrm{d}V$. In differential-geometric terms, this means that we have a smooth, nowhere-vanishing, totally skew tensor field $\epsilon_{a_1 \cdots a_n}$ on M. Our Hilbert space, and operators, are now defined in terms of certain fields on M.

We first define the Hilbert space. Consider the collection H of all complex-valued, measurable, square-integrable functions f on M. This H is certainly a complex vector apace. We introduce a norm on H:

$$\|f\|^2 = \int_M f f^* \, \mathrm{d}V \tag{26}$$

It is known that this H thus becomes a Hilbert space. (Actually, we have been a little sloppy here. One should, more properly, define an equivalence relation on H: two functions are equivalent if they differ only on a subset (of M) of measure zero. It is the equivalence classes which actually form a Hilbert space. For example, the function f which vanishes everywhere on M except one point, where it is one, is measurable and square-integrable. Its norm, (26), is zero, although this f is not the zero element of H. It is, however, in the zero equivalence class, for it differs from the zero function only on a set (namely, one point) of measure zero.) This is a special case of a more general theorem: the

collection of all complex-valued, measurable, square-integrable functions (more precisely, the collection of equivalence classes as above) on a complete measure space form a Hilbert space.

We now introduce some operators. Let v^a be any smooth (complex) contravariant vector field, and v any smooth (complex) scalar field on M. Then with each smooth, complex-valued function f on M we may associate the function

$$Vf = v^a \nabla_a f + vf \tag{27}$$

where ∇_a denotes the gradient on M. To what extent does (27) define an operator on H? Unfortunately, (27) is not applicable to every element of H, for two reasons: i) a function f could be measurable and square-integrable (i.e., an element of H), but not differentiable. Then the gradient operation in (27) would not be defined. ii) an element f of H could even be smooth, but could have the property that, although f itself is square-integrable, the function (27) is not. However, there is a large class of elements of H on which (27) is defined and results in an element of H. Such a class, for example, is the collection of all functions F which are smooth and have compact support. (Such a function is automatically square-integrable and measurable.) This class is, in fact, dense in H. Clearly, (27) is linear whenever it is defined. Thus, we can call (27) an "operator on H", in the sense that we have agreed to abuse that term.

We agree to call an operator *Hermitian* if, whenever Vf and Vg are defined, $(Vf, g) = (f, Vg)$. What are the necessary and sufficient conditions that (27) be Hermitian? Let f and g be smooth functions on M, of compact support. Then:

$$\begin{aligned}(Vf, g) &= \int_M (v^a \nabla_a f + vf) g^* \, \mathrm{d}V \\ &= \int_M [-f v^a \nabla_a g^* + f g^* (-\nabla_a v^a + v)] \, \mathrm{d}V\end{aligned} \tag{28}$$

where we have done an integration by parts (throwing away a surface term by the compact supports). Eqn. (28) is clearly equal to

$$(f, Vg) = \int_M [f v^{*a} \nabla_a g^* + f v^* g^*] \, \mathrm{d}V \tag{29}$$

for every f and g when and only when:

$$v^{*a} = -v^a \qquad v - v^* = \nabla_a v^a \tag{30}$$

These, then, are the necessary and sufficient conditions that V be Hermitian. One further remark is required with regard to what the divergence in the first Eqn. (30) is supposed to mean. (We don't have a metric, or a covariant derivative, defined on M.) It is well-known that the divergence of a contravariant vector field can be defined on a manifold with a volume-element $\epsilon_{a_1 \cdots a_n}$. This

can be done, for example, using either exterior derivatives or Lie derivatives. For instance, using Lie derivatives we define "$\nabla_a v^a$" by:

$$\mathcal{L}_{v^m}\epsilon_{a_1\cdots a_n} = (\nabla_a v^a)\epsilon_{a_1\cdots a_n} \tag{31}$$

(Note that, since the left side is totally skew, it must be *some* multiple of $\epsilon_{a_1\cdots a_n}$.)

Finally, we work out the commutator of two of our operators; $V = (v^a, v)$ and $W = (w^a, w)$. If f is a smooth function of compact support, we have:

$$\begin{aligned}[V,W]f &= (v^a\nabla_a + v)(w^b\nabla_b + w)f - (w^b\nabla_b + w)(v^a\nabla_a + v)f \\ &= (v^b\nabla_b w^a - w^b\nabla_b v^a)\nabla_a f + (v^a\nabla_a w - w^a\nabla_a v)f\end{aligned} \tag{32}$$

Note that the commutator is again an operator of the form we have been discussing, (27). Note furthermore that the vector part of the commutator is the Lie bracket of the vector fields appearing in V and W.

To summarize, with any n-manifold M on which there is given a smooth, nowhere-vanishing volume element we associate a Hilbert space H along with a collection of operators on H. The commutator of two operators in this collection is again an operator in the collection.

5 Hilbert Space and Operators for the Klein-Gordon Equation

We now complete our description of the quantum theory of a free, relativistic, spin-zero particle.

For our Hilbert space we take, as suggested by Sec. 3, the collection of all complex-valued, measurable, square-integrable functions on the future mass shell, M_μ^+. In order to obtain position, momentum, energy, etc. operators, we use the scheme described in Sec. 4. That is, we look for vector and scalar fields on M_μ^+.

We first consider momentum operators. Let p^a be any constant vector field in Minkowski apace, and ϕ any positive-frequency solution of the Klein-Gordon equation. Then, clearly,

$$\frac{\hbar}{i}p^a\nabla_a\phi \tag{33}$$

is also a positive-frequency solution of the Klein-Gordon equation. In terms of the corresponding functions on M_μ^+, (33) takes the form

$$f \to (\hbar p^a k_a)f \tag{34}$$

That is to say, we multiply f by the real function $(\hbar p^a k_a)$ on M_μ^+. Thus, for each constant vector field p^a, we have an operator, $P(p^a)$, on our Hilbert space H. Since the multiplying function in (34) is real, the operators $P(p^a)$ are all Hermitian. (See (30).) We now interpret these operators. Choose a constant,

unit, future-directed timelike vector field t^a in Minkowski space (a preferred "state of rest"). Then $P(t^a)$ is the "energy" operator, and $P(p^a)$, with p^a unit and perpendicular to t^a, is the "component of momentum in the p^a-direction" operator.

The position operators are more complicated. Not only do they depend on more objects in Minkowski space (rather than just a single p^a as in the momentum case), but also they require us to take derivatives in the mass shell. To obtain a position operator, we need the following information: a choice of origin O in Minkowski space, a constant, unit, future-directed timelike vector field t^a in Minkowski space, and a constant unit vector field q^a which is perpendicular to t^a. (Roughly speaking, O and t^a define a spacelike 3-plane — the "instant" at which the operator is to be applied — q^a defines "which position coordinate we're operating with", and O tells us what the origin of this position coordinate is.) Now, q^a is a vector in momentum space, and therefore defines a constant vector field in momentum space, which we also write as q^a. One is tempted to take the derivative of f along this vector field. But this will not work, for q^a is not tangent to the mass shell, whereas f is only defined on the mass shell. To correct this deficiency, we project q^a into the mass shell — that is, we add to q^a that multiple of t^a which results in a vector field lying in M_μ^+:

$$-\frac{1}{i}\left[q^a - t^a(t^b k_b)^{-1}(q^c k_c)\right] \tag{35}$$

We now have a vector field on M_μ^+, and therefore an operator on our Hilbert space H. But are those operators Hermitian? From (30), we see that this question reduces to the question of whether the divergence of (35) vanishes or not. Unfortunately, we obtain for this divergence

$$-\frac{1}{i}\left(g^{ab} - \mu^{-2}k^a k^b\right)\partial_a\left[q^b - t^b(t^c k_c)^{-1}(q^d k_d)\right] = -\frac{1}{i}(q^a k_a)(t^b k_b)^{-2} \tag{36}$$

where we have denoted the derivative in momentum space by ∂_a. To obtain a Hermitian operator, we take the Hermitian part of the operator represented by (35):

$$f \to -\frac{1}{i}\left[q^a - t^a(t^b k_b)^{-1}(q^c k_c)\right]\partial_a f - \frac{1}{2i}(q^a k_a)(t^b k_b)^{-2} \tag{37}$$

In (37), f is to be the function on M_μ^+ obtained using O as the origin (see (18).) (Why is there no $\hbar$ in (37)? We should, perhaps, have called k-space "wave number and frequency space" rather than "momentum space".) We shall write the operator (37) $X(O, t^a, q^a)$. For any value of its arguments, X is a Hermitian operator on H. (It is strange — and perhaps rather unpleasant — that the position and momentum operators are so different from each other.)

We now have a lot of operators, and so we can ask for their commutators. This is easily done by substituting into our explicit formula, Eqn. (32). The

result is the standard formulae:

$$\begin{aligned}
&\left[P(p^a), P(p'^a)\right] = 0 \\
&\left[X(O,t^a,q^a), X(O,t^a,q'^a)\right] = 0 \\
&[P(p^a), X(O,t^a,q^a)] = -i\hbar(p^a q_a) \\
&(p^a t_a) = 0
\end{aligned} \tag{38}$$

The next thing one normally does with operators (in the Heisenberg representation, which is the one we're using) is to work out their time derivatives. For the momentum operators, this is easy, for no notion of a "time" was used to define $P(p^a)$). Thus, whatever reasonable thing one wants to mean by a " $\dot{}$ ", we have:

$$\dot{P}(p^a) = 0 \tag{39}$$

This, of course, is what we expect for the momentum operator on a free particle. For the position operators, on the other hand, we have an interesting notion of time-derivative. We want to compare $X(O,t^a,q^a)$ with "the same position operator at a slightly later time". This "at a slightly later time" is expressed by slightly displacing O in the t^a-direction. Thus, we are led to define:

$$\dot{X}(O,t^a,q^a) = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[X(O',t^a,q^a) - X(O,t^a,q^a)\right] \tag{40}$$

where O' is defined by the property that its position vector relative to O is ϵt^a. It is straightforward to check, with this definition, that

$$\dot{X}(O,t^a,q^a)f = -(q^a k_a)(t^b k_b)^{-1} f \tag{41}$$

which, of course, is what we expected. Note that a number of statements about how $X(O,t^a,q^a)$ depends on its arguments follow directly from Eqn. (41).

Finally, one would like to ask about the eigenvectors and eigenvalues of our operators. It is clear from Eqn. (34) that the only candidate for an eigenfunction of $P(p^a)$ would be a δ-function on M_μ^+. Of course, a δ-function is not a function, and hence not an element of H (we cannot enlarge H to include such functions, if we want to keep a Hilbert space, for a δ-function should not be square-integrable.) It is convenient to have the idea, however, that if $P(p^a)$ had eigenfunctions, they would be plane-waves. We next ask for eigenfunctions of $X(O,t^a,q^a)$. We look for the wave function of a "particle localized at the origin", that is we look for an f such that $X(O,t^a,q^a)f = 0$ for *every* q^a which is perpendicular to t^a (t^a and O fixed). That is, from (37), we require that

$$\left[q^a - t^a(t^b k_b)^{-1}(q^c k_c)\right]\partial_a f - \frac{1}{2}(q^a k_a)(t^b k_b)^{-2} f = 0 \tag{42}$$

for every such q^a. The solution to (42) is:

$$f = \text{const.}\,(t^a k_a)^{1/2} \tag{43}$$

The first remark concerning (43) is that it is not square-integrable, and hence does not represent an element of H. This does not stop us, however, from

substituting (43) into (13) to obtain a function ϕ on Minkowski space. The resulting ϕ (the explicit formula is not very enlightening — it involves Hankel functions) is not a δ-function at O. In fact, this ϕ is "spread out" around O to distances of the order of μ^{-1} — the Compton wavelength of our particle. Thus, our picture is that a relativistic particle cannot be confined to distances much smaller than its Compton wavelength.

6 The Direct Sum of Hilbert Spaces

Associated with any countable sequence, $H', H'', H''', \ldots$, of Hilbert spaces there is another Hilbert space, written $H' \oplus H'' \oplus H''' \oplus \ldots$, and called the *direct sum* of $H', H'', H''', \ldots$. We shall give the definition of the direct sum and a few of its elementary properties.

Consider the collection of all sequences

$$(\xi', \xi'', \xi''', \ldots) \tag{44}$$

consisting of one element (ξ') of H', one element (ξ'') of H'', etc., for which the sum

$$\|\xi'\|^2 + \|\xi''\|^2 + \|\xi''\|^2 + \ldots \tag{45}$$

converges. This collection is the underlying point set of the direct sum. To obtain a Hilbert space, we must define addition, scalar multiplication, and an inner product, and verify **H1**–**H11**.

The sum of two sequences (44) is defined by adding them "component-wise":

$$(\xi', \xi'', \xi''', \ldots) + (\eta', \eta'', \eta''', \ldots) = (\xi' + \eta', \xi'' + \eta'', \xi''' + \eta''', \ldots) \tag{46}$$

We must verify that, if the addends satisfy (45), then so does the sum. This follows immediately from the inequality:

$$\begin{aligned} \|\xi' + \eta'\|^2 &= \|\xi'\|^2 + (\xi', \eta') + (\eta', \xi') + \|\eta'\|^2 \\ &\leq \|\xi'\|^2 + 2\|\xi'\|\|\eta'\| + \|\eta'\|^2 \\ &\leq 2\|\xi'\|^2 + 2\|\eta'\|^2 \end{aligned} \tag{47}$$

The product of a sequence (44) and a complex number α is defined by:

$$\alpha(\xi', \xi'', \xi''', \ldots) = (\alpha\xi', \alpha\xi'', \alpha\xi''', \ldots) \tag{48}$$

That the right side of (48) satisfies (45) follows from the fact that

$$\|\alpha\xi'\| = |\alpha|\|\xi'\| \tag{49}$$

We have now defined addition and scalar multiplication. That these two operations satisfy **H1**–**H7**, i.e., that we have a complex vector space, is trivial.

We define the inner product between two sequences (44) to be the complex number

$$((\xi', \xi'', \xi''', \ldots), (\eta', \eta'', \eta''', \ldots)) = (\xi', \eta') + (\xi'', \eta'') + (\xi'', \eta'') + \ldots \tag{50}$$

The indicated sum of complex numbers on the right of (50) converges if (in fart, converges absolutely if and only if) the sum of the absolute values converges. Thus, the absolute convergence of the right side of (50) follows from the fact that

$$|(\xi', \eta')| \leq \|\xi'\| \leq \frac{1}{2}\|\xi'\|^2 + \frac{1}{2}\|\eta'\|^2 \tag{51}$$

We now have a complex vector space in which there is defined an inner product. (Note incidentally, that the norm is given by (45).) The verification of **H8**, **H9**, and **H10** is easy.

Thus, as usual, the only difficult part is to check **H11**. Consider a Cauchy sequence of sequences (44). That is to say, we have a countable collection of such sequences,

$$\begin{aligned} \phi_1 &= (\xi_1', \xi_1'', \xi_1''', \ldots) \\ \phi_2 &= (\xi_2', \xi_2'', \xi_2''', \ldots) \\ \phi_3 &= (\xi_3', \xi_3'', \xi_3''', \ldots) \\ &\vdots \end{aligned} \tag{52}$$

with the following property: for each real $\epsilon > 0$ there is a number N such that

$$\|\phi_i - \phi_j\|^2 = \|\xi_i' - \xi_j'\|^2 + \|\xi_i'' - \xi_j''\|^2 + \cdots \leq \epsilon \tag{53}$$

whenever $i, j \geq N$. We must show that the sequence of elements (52) of the direct sum converge to some element of the direct sum. First note that (53) implies

$$\|\xi_i' - \xi_j'\|^2 \leq \epsilon, \quad \|\xi_i'' - \xi_j''\|^2 \leq \epsilon, \quad \ldots \tag{54}$$

That is to say, the first "column" of (52) is a Cauchy sequence in H', the second column a Cauchy sequence in H'', etc. Since $H', H'', \ldots$ are Hilbert spaces, these Cauchy sequences converge, say, to $\xi' \in H'$, $\xi'' \in H''$, etc. Form

$$\phi = (\xi', \xi'', \xi''', \ldots) \tag{55}$$

We must show that the ϕ_i converge to ϕ, and that is ϕ is an element of the direct sum (i.e., that (45) converges for ϕ). Fix $\epsilon \geq 0$ and choose i such that $\|\phi_i - \phi_j\|^2 \leq \epsilon$ whenever $j > i$. Then, for each positive integer n,

$$\|\xi_i' - \xi_j'\|^2 + \|\xi_i'' - \xi_j''\|^2 + \ldots + \|\xi_i^{(n)} - \xi_j^{(n)}\|^2 \leq \epsilon \tag{56}$$

Taking the limit of (56) as $j \to \infty$, we obtain

$$\|\xi_i' - \xi'\|^2 + \|\xi_i'' - \xi''\|^2 + \ldots + \|\xi_i^{(n)} - \xi^{(n)}\|^2 \leq \epsilon \tag{57}$$

but n is arbitrary, and so, taking the limit of (57) as $n \to \infty$,

$$\|\phi_i - \phi\|^2 = \|\xi_i' - \xi'\|^2 + \|\xi_i'' - \xi''\|^2 + \ldots \leq \epsilon \tag{58}$$

That is to say, the ϕ_i converge to ϕ. Finally, the fact that ϕ is actually an element of the direct sum, i.e., the fact that

$$\|\xi'\|^2 + \|\xi''\|^2 + \|\xi'''\|^2 + \dots \tag{59}$$

converges, follows immediately by substituting

$$\|\xi'\|^2 \leq 2\|\xi_i'\|^2 + 2\|\xi_i' - \xi'\|^2 \tag{60}$$

(and the corresponding expressions with more primes) into (59), and using (58) and the fact that ϕ_i is an element of the direct sum. Thus, the direct sum is complete.

To summarize, we have shown how to construct a Hilbert space from a countable sequence of Hilbert spaces. Note, incidentally, that the direct sum is essentially independent of the order in which the Hilbert spaces are taken. More precisely, the direct sum obtained by taking $H', H'', \dots$ in one order is naturally isomorphic to the direct sum obtained by taking these spaces in any other order.

Finally, we discuss certain operators on the direct sum. Consider a sequence of operators: A' acting on H', A'' acting on H'', etc. Then with each element (44) of the direct sum we may associate the sequence

$$(A'\xi', A''\xi'', A'''\xi''', \dots) \tag{61}$$

Unfortunately, (61) may not be an element of the direct sum, for

$$\|A'\xi'\|^2 + \|A''\xi''\|^2 + \|A'''\xi'''\|^2 + \dots \tag{62}$$

may fail to converge. However, (61) will produce an element of the direct sum when acting on a certain dense subset of the direct sum, namely, the set of sequences (44) which consist of zeros after a certain point. Is there any condition on the A's which will ensure that (61) will always be an element of the direct sum? An operator A on a Hilbert space H is said to be *bounded* if A is defined everywhere and, for some number a, $\|A\xi\| \leq a\|\xi\|$ for every $\xi \in H$. The smallest such a is called the *bound* of A, written $|A|$. (The norm on a Hilbert space induces on it a metric topology. Boundedness is equivalent to continuity in this topology.) It is clear that (62) converges for every element of the direct sum provided i) all the A's are bounded, and ii) the sequence of real numbers $|A'|, |A''|, \dots$ is bounded.

7 The Completion of an Inner-Product Space

It is sometimes the case, when one wishes to construct a Hilbert space, that one finds a set on which addition, scalar multiplication, and an inner product are defined, subject to **H1**–**H10** — what we shall call an inner product space. One wants, however, to obtain a Hilbert space, i.e., something which also satisfies

H11. There is a construction for obtaining a Hilbert space from an inner product space. Since this construction is in most textbooks, we merely indicate the general idea.

Let G be an inner product space. Denote by G' the collection of all Cauchy sequences in G'. If $\xi_i \in G', \eta_i \in G'$ $(i = 1, 2, \ldots)$, we write $\xi_i \approx \eta_i$ provided

$$\lim_{i\to\infty} \|\xi_i - \eta_i\| = 0 \tag{63}$$

One verifies that "$\approx$" is an equivalence relation. The collection of equivalence classes, denoted by $\overline{G}$, is to be made into a Hilbert space.

Consider two elements of $\overline{G}$, i.e., two equivalence classes, and let ξ_i and η_i be representatives. We define a new sequence, whose ith element is $\xi_i + \eta_i$. One verifies, using the fact that ξ_i and η_i, are Cauchy sequences, that this new sequence is also Cauchy. Furthermore, if ξ_i and η_i are replaced by equivalent Cauchy sequences, the sum becomes a Cauchy sequence which is equivalent to $\xi_i + \eta_i$. Thus, we have defined an operation of addition in $\overline{G}$. In addition, if ξ_i is a Cauchy sequence and α a complex number, $\alpha\xi_i$ is a Cauchy sequence whose equivalence class depends only on the equivalence class of ξ_i. We have thus defined an operation of scalar multiplication in $\overline{G}$. These two operations satisfy **H1**–**H7**.

If $\xi_i \in G'$, $\eta_i \in G'$, then

$$\lim_{i\to\infty} (\xi_i, \eta_i) \tag{64}$$

exists. Furthermore, this complex number is unchanged if ξ_i and η_i are replaced by equivalent Cauchy sequences. Thus, (64) defines an inner product on $\overline{G}$. One must now verify **H8**, **H9**, and **H10**, so that $\overline{G}$ becomes an inner product space. Finally (and this is the only hard part), one proves that $\overline{G}$ is complete, and so constitutes a Hilbert space. The Hilbert space $\overline{G}$ is called the *completion* of the inner product space G.

Note that G can be considered as a subspace of its completion; with each $\xi \in G$ associate the element of $\overline{G}$ (the equivalence class) containing the Cauchy sequence $\xi_i = \xi$ of G. It is easily checked from the definition, in fact, that G is dense in $\overline{G}$. Suppose that G itself were already complete? Then every Cauchy sequence in G would converge, and, from (63), two Cauchy sequences would be equivalent if and only if they converged to the same thing. Thus, in this case $\overline{G}$ would be just another copy of G; the completion of a complete space is just that space again.

Finally, suppose that A is a bounded operator on the inner product space G. Then A can be extended to a bounded operator on $\overline{G}$. That is, there is a bounded operator on $\overline{G}$ which reduces to A on G considered as a subspace of $\overline{G}$. To prove this, let ξ_i be a Cauchy sequence in G. Then, since A is bounded,

$$\|A\xi_i - A\xi_j\| = \|A(\xi_i - \xi_j)\| \leq |A|\|\xi_i - \xi_j\| \tag{65}$$

whence $A\xi_i$ is a Cauchy sequence in G. Furthermore, if two Cauchy sequences satisfy (63), then

$$\lim_{i\to\infty} \|A\xi_i - A\eta_i\| \leq |A| \lim_{i\to\infty} \|\xi_i - \eta_i\| = 0 \tag{66}$$

That is to say, $A\xi_i$ is replaced by an equivalent Cauchy sequence when ξ_i is. Therefore, we can consider A as acting on elements of $\overrightarrow{G}$ to produce elements of $\overrightarrow{G}$. This action is clearly linear, and so we have an operator $\overrightarrow{A}$ defined on $\overrightarrow{G}$. Furthermore, this $\overrightarrow{A}$ is bounded, and in fact $|\overrightarrow{A}| = |A|$, for

$$\lim_{i\to\infty} \|A\xi_i\| \leq |A| \lim_{i\to\infty} \|\xi_i\| \tag{67}$$

for any Cauchy sequence in G.

8 The Complex-Conjugate Space of a Hilbert Space

Let H be a Hilbert space. We introduce the notion of the *complex-conjugate space* of H, written $\bar{H}$. As point-sets, $H = \bar{H}$. That is to say, with each element $\xi \in H$ there is associated an element of $\bar{H}$; this element will be written $\bar{\xi}$. Furthermore, we take as the group structure on $\bar{H}$ that is induced from H:

$$\bar{\xi} + \bar{\eta} = \overline{(\xi + \eta)} \tag{68}$$

In other words, the sum of two elements of $\bar{H}$ is defined by taking the sum (in H) of the corresponding elements of H, and taking the result back to $\bar{H}$. Scalar multiplication in H, on the other hand, is defined by the formula ($\mu \in \mathbb{C}$, $\bar{\xi} \in \bar{H}$):

$$\mu\bar{\xi} = \overline{(\bar{\mu}\xi)} \tag{69}$$

That is, to multiply an element of $\bar{H}$ by a complex number, one multiplies the corresponding element of H by the complex-conjugate of that number, and takes the result back to $\bar{H}$. (Note that a bar appears in two different senses in (69). A bar over a complex number denotes its complex-conjugate; a bar over an element of H denotes the corresponding element of $\bar{H}$.) Finally, the inner product on $\bar{H}$ is fixed by requiring that the transition from H to $\bar{H}$ preserve norms:

$$\|\bar{\xi}\| = \|\xi\| \tag{70}$$

It is obvious that this $\bar{H}$ thus becomes a Hilbert space.

Note that the complex-conjugate space of $\bar{H}$ is naturally isomorphic with H. We write $\bar{\bar{H}} = H$, and, for $\xi \in H$, $\bar{\bar{\xi}} = \xi$.

The reason for introducing $\bar{H}$ is that one frequently encounters mappings on H which are anti-linear ($T(\mu\xi + \eta) = \bar{\mu}T(\xi) + T(\eta)$) rather than linear ($T(\mu\xi+\eta) = \mu T(\xi)+T(\eta)$). Anti-linear mappings on H become linear mappings on $\bar{H}$, and it is easier to think about linear mappings than anti-linear ones. Consider, for example, the inner product on H, (ξ, η). This can be considered as a mapping $H \times H \to \mathbb{C}$, which is linear in the first H and anti-linear in the second. If, however, we consider the inner product as a mapping $H \times \bar{H} \to \mathbb{C}$, it becomes linear in both factors.

9 The Tensor Product of Hilbert Spaces

With any finite collection of Hilbert spaces, $H^\alpha, H^\beta, \ldots, H^\gamma$, there is associated another Hilbert space, called the *tensor product* of $H^\alpha, H^\beta, \ldots, H^\gamma$, and written $H^\alpha \otimes H^\beta \otimes \cdots \otimes H^\gamma$. We shell define the tensor product and derive a few of its properties.

It is convenient to introduce an index notation. We attach a raised Greek index to a vector to indicate to which Hilbert space it belongs, e.g., $\xi^\alpha \in H^\alpha$, $\eta^\beta \in H^\beta$. The corresponding complex-conjugate spaces, $\overline{H^\alpha}, \overline{H^\beta}, \ldots, \overline{H^\gamma}$, will be written $\bar{H}_\alpha, \bar{H}_\beta, \ldots, \bar{H}_\gamma$. Membership in the complex-conjugate spaces will be indicated with a lowered Greek index. The element of $\bar{H}_\alpha$ which corresponds to $\xi^\alpha \in H^\alpha$ would be written $\bar{\xi}_\alpha$, while (69) would be written thus: $\mu\bar{\xi}_\alpha = \overline{(\bar{\mu}\,\xi^\alpha)}$. Finally, the operation of taking the inner product (which associates a complex number, linearly, with an element of H and an element of $\bar{H}$) is indicated by placing the two elements next to each other, e.g., $\xi^\alpha\bar{\eta}_\alpha$. Hence, $\|\xi^\alpha\|^2 = \xi^\alpha\bar{\xi}_\alpha$. The inner product operation looks (and is) similar to "contraction."

We now wish to define the tensor product. In order to avoid cumbersome strings of dots, we shall discuss the tensor product of just two Hilbert spaces, H^α and H^β. The tensor product of n Hilbert spaces is easily obtained by inserting dots at appropriate places in the discussion below.

Consider the collection of all formal expressions of the following type:

$$\xi^\alpha\eta^\beta + \ldots + \sigma^\alpha\tau^\beta \tag{71}$$

"Formal" here means that the pluses and juxtapositions of elements in (71) are not to be considered, for the moment, as well-defined operations. They are merely marks on the paper. We introduce, on the collection of all such formal sums, an equivalence relation: two formal sums will be considered equivalent if they can be obtained from each other by any combination of the following operations on such sums:

1. Permute, in any way, the terms of a formal sum.
2. Add to a formal sum, or delete, the following combination of terms: $(\mu\xi^\alpha)\eta^\beta + (-\xi^\alpha)(\mu^{-1}\eta^\beta)$.
3. Add to a formal sum, or delete, the following combination of terms: $(\xi^\alpha + \sigma^\alpha)\eta^\beta + (-\xi^\alpha)\eta^\beta + (-\sigma^\alpha)\eta^\beta$.
4. Add to a formal sum, or delete, the following combination of terms: $\xi^\alpha(\eta^\beta + \tau^\beta) + \xi^\alpha(-\eta^\beta) + \xi^\alpha(-\tau^\beta)$.

We denote the collection of equivalence classes by $F^{\alpha\beta}$. The idea is to introduce on this $F^{\alpha\beta}$ the structure of an inner-product space, and then take the completion to obtain the tensor product.

We add two formal sums by "stringing them together":

$$(\xi^\alpha\eta^\beta+\ldots+\sigma^\alpha\tau^\beta)+(\phi^\alpha\psi^\beta+\ldots+\lambda^\alpha\kappa^\beta) = \xi^\alpha\eta^\beta+\ldots+\sigma^\alpha\tau^\beta+\phi^\alpha\psi^\beta+\ldots+\lambda^\alpha\kappa^\beta \tag{72}$$

The equivalence class of the formal sum on the right in (72) depends, of course, only on the equivalence classes of the two formal sums on the left, and so we have defined an operation of addition on $F^{\alpha\beta}$. Similarly, the product of a format sum and a complex number, defined by

$$\mu(\xi^\alpha\eta^\beta + \ldots + \sigma^\alpha\tau^\beta) = (\mu\xi^\alpha)\eta^\beta + \ldots + (\mu\sigma^\alpha)\tau^\beta \tag{73}$$

induces an operation of scalar multiplication on $F^{\alpha\beta}$. (Note that we don't change the equivalence class of the right side of (73) by placing some or all of the μ's with the second vectors rather than the first.) Thus, we have on $F^{\alpha\beta}$ the structure of a complex vector space. So far, we have merely repeated the standard construction of the tensor product of two vector spaces.

We next wish to define an inner product, or, equivalently, a norm, on $F^{\alpha\beta}$. The norm of a formal sum, Eqn. (71), is defined by writing

$$(\xi^\alpha\eta^\beta + \cdots + \sigma^\alpha\tau^\beta)(\bar{\xi}_\alpha\bar{\eta}_\beta + \cdots + \bar{\sigma}_\alpha\bar{\tau}_\beta) \tag{74}$$

and expanding using associativity. For example, the norm of a formal sum with just two terms would be given by the sum of the complex numbers on the right of:

$$\begin{aligned}(\xi^\alpha\eta^\beta + \sigma^\alpha\tau^\beta)(\bar{\xi}_\alpha\bar{\eta}_\beta + \bar{\sigma}_\alpha\bar{\tau}_\beta)& \\ = (\xi^\alpha\bar{\xi}_\alpha)(\eta^\beta\bar{\eta}_\beta) + (\sigma_\alpha\bar{\xi}_\alpha)(\tau^\beta\bar{\eta}_\beta) + (\xi^\alpha\bar{\sigma}_\alpha)(\eta^\beta\bar{\tau}_\beta) + (\sigma^\alpha\bar{\sigma}_\alpha)(\tau^\beta\bar{\tau}_\beta)&\end{aligned} \tag{75}$$

This norm clearly depends only on the equivalence class of the formal sum, and so defines a norm, and hence an inner product on $F^{\alpha\beta}$. This inner product on $F^{\alpha\beta}$ certainly satisfies **H8** and **H9** — but does it satisfy also **H10**? To show that it does involves a bit more work. A formal sum will be said to be in *normal form* if any two elements of H^α appearing in that sum are either proportional to each other (parallel) or have vanishing inner product with each other (perpendicular) and if, furthermore, any two elements of H^β appearing in that sum are also either parallel or perpendicular, and if, finally, no two terms in that sum have the property that both their H^α elements and their H^α elements are parallel. (This last condition can always be achieved by combining terms.) The norm of a formal sum in normal form is clearly positive. For example, if Eqn. (71) were in normal form, its norm would be

$$\left\|\xi^\alpha\eta^\beta + \ldots + \sigma^\alpha\tau^\beta\right\|^2 = (\xi^\alpha\bar{\xi}_\alpha)(\eta^\beta\bar{\eta}_\beta) + \ldots + (\sigma^\alpha\bar{\sigma}_\alpha)(\tau^\beta\bar{\tau}_\beta) \tag{76}$$

Thus, the proof that the norm we have defined on $F^{\alpha\beta}$ is positive-definite will be complete if we can show that every formal sum is equivalent to a formal sum in normal form. The essential step in this demonstration is the Gram-Schmidt orthogonalization procedure. Let $\xi, \eta_1, \ldots, \eta_n$ be elements of a Hilbert space H. Then

$$\xi = \mu_1\eta_1 + \cdots + \mu_n\eta_n + \tau \qquad\qquad (\mu_i \in \mathbb{C}, (\eta_i, \tau) = 0) \tag{77}$$

That is to say, any vector in H can be written as a linear combination of $\eta_1, \ldots, \eta_n$, plus a vector perpendicular to the η's. Consider now a formal sum, say

$$\xi^\alpha\eta^\beta + \sigma^\alpha\tau^\beta + \phi^\alpha\psi^\beta \tag{78}$$

We obtain an equivalent formal sum by replacing σ^α by a vector parallel to ξ^α plus a vector perpendicular to ξ^α and combining terms. Thus, Eqn. (78) is equivalent to a formal sum

$$\xi^\alpha \eta'^\beta + \sigma'^\alpha \tau^\beta + \phi^\alpha \psi^\beta \tag{79}$$

in which σ'^α is perpendicular to ξ^α. We next obtain an equivalent formal sum by replacing ϕ^α by a vector parallel to σ'^α plus a vector parallel to ξ^α plus a vector perpendicular to both σ'^α and ξ^α. Thus, Eqn. (79) is equivalent to a formal sum

$$\xi^\alpha \eta''^\beta + \sigma'^\alpha \tau'^\beta + \phi'^\alpha \psi^\beta \tag{80}$$

with ξ^α, σ'^α, and ϕ'^α all perpendicular to each other. We now repeat this procedure with the H^β vectors to obtain a formal sum in normal form which is equivalent to Eqn. (78). Hence, every formal sum is equivalent to a formal sum in normal form, whence the norm on $F^{\alpha\beta}$ is positive-definite, whence $F^{\alpha\beta}$ is an inner-product space.

We now define the *tensor product* of H^α and H^β to be the completion of $F^{\alpha\beta}$:

$$H^\alpha \otimes H^\beta = \overline{F^{\alpha\beta}} \tag{81}$$

This is really quite complicated. An element of the tensor product is an equivalence class of Cauchy sequences in an inner-product space whose elements are equivalence classes of formal sums. Note that $F^{\alpha\beta}$ itself can be considered as a subspace of $H^\alpha \otimes H^\beta$. Elements of the tensor product which belong to $F^{\alpha\beta}$ will be called *finite* elements. In fact, we shall go one step further and consider formal sums (71) to be elements of the tensor product. Equivalent formal sums are then equal as elements of the tensor product. With these conventions, we shall be able to avoid, for the most part, having to speak always in terms of equivalence classes and Cauchy sequences.

The tensor product of more than two Hilbert spaces, $H^\alpha \otimes H^\beta \otimes \cdots \otimes H^\gamma$, is defined in a completely analogous way. We use Greek indices to indicate membership in the various tensor products, e.g., we write $\xi^{\alpha\beta\cdots\gamma}$ for a typical element of $H^\alpha \otimes H^\beta \otimes \cdots \otimes H^\gamma$. Addition and scalar multiplication within the tensor products is indicated in the obvious way:

$$\begin{aligned} &\xi^{\alpha\beta\cdots\gamma} + \eta^{\alpha\beta\cdots\gamma} \\ &\mu\xi^{\alpha\beta\cdots\gamma} \end{aligned} \tag{82}$$

Note, incidentally, that our original formal sums are now considerably less formal. For example, Eqn. (71) can be considered as the sum, in $H^\alpha \otimes H^\beta$, of the following elements of $H^\alpha \otimes H^\beta$: $\xi^\alpha \eta^\beta, \ldots, \sigma^\alpha \tau^\beta$.

We next observe that there is a natural, one-to-one correspondence between the formal sums which are used to obtain $H^\alpha \otimes \cdots \otimes H^\beta$,

$$\xi^\alpha \cdots \eta^\beta + \cdots + \sigma^\alpha \cdots \tau^\beta \tag{83}$$

and those which are used to obtain $\bar{H}_\alpha \otimes \cdots \otimes \bar{H}_\beta$:

$$\bar{\xi}_\alpha \cdots \bar{\eta}_\beta + \cdots + \bar{\sigma}_\alpha \cdots \bar{\tau}^\beta \tag{84}$$

That is to say, the inner product space consisting of finite elements of $\bar{H}_\alpha \otimes \cdots \otimes \bar{H}_\beta$ is the complex conjugate space of the inner product space consisting of finite elements of $H^\alpha \otimes \cdots \otimes H^\beta$. This relationship clearly continues to hold in the actual tensor product spaces. We conclude that $\bar{H}_\alpha \otimes \cdots \otimes \bar{H}_\beta$ is the same as (words we shall frequently use instead of "is naturally isomorphic with") $\overline{H^\alpha \otimes \cdots \otimes H^\beta}$. The tensor product of the complex-conjugate spaces is the same as the complex-conjugate space of the tensor product. This fact allows us to extend our index notation still further. A typical element of $\bar{H}_\alpha \otimes \cdots \otimes \bar{H}_\beta$ will be written $\bar{\xi}_{\alpha\cdots\beta}$, this being the element which corresponds to the element $\xi^{\alpha\cdots\beta}$ of $H^\alpha \otimes \cdots \otimes H^\beta$. The inner product, in the tensor product space, of two elements, $\xi^{\alpha\cdots\beta}$ and $\eta^{\alpha\cdots\beta}$, of $H^\alpha \otimes \cdots \otimes H^\beta$ can now be written as follows:

$$\xi^{\alpha\cdots\beta}\bar{\eta}_{\alpha\cdots\beta} \tag{85}$$

Thus, the index notation extends very nicely from the original collection of Hilbert spaces to the various tensor products which can be constructed.

We now introduce some operations between elements of our tensor product spaces. Let $\xi^{\alpha\cdots\beta\gamma\cdots\delta}$ and $\bar{\eta}_{\gamma\ldots\delta}$ be finite elements of $H^\alpha \otimes \ldots \otimes H^\beta \otimes H^\gamma \otimes \ldots \otimes H^\gamma$ and $\bar{H}_\gamma \otimes \ldots \otimes \bar{H}_\delta$, respectively. We can certainly associate with these two an element, $\xi^{\alpha\cdots\beta\gamma\cdots\delta}\bar{\eta}_{\gamma\ldots\delta}$ of $H^\alpha \otimes \ldots \otimes H^\beta$. For example,

$$\begin{aligned}
&(\xi^\alpha\eta^\beta\sigma^\gamma\tau^\delta + \phi^\alpha\psi^\beta\lambda^\gamma\kappa^\delta)(\bar{\rho}_\gamma\bar{\epsilon}_\delta + \bar{\nu}_\gamma\bar{\zeta}_\delta) \\
&\quad = \xi^\alpha\eta^\beta(\sigma^\gamma\bar{\rho}_\gamma)(\tau^\delta\bar{\epsilon}_\delta) + \phi^\alpha\psi^\beta(\lambda^\gamma\bar{\rho}_\gamma)(\kappa^\delta\bar{\epsilon}_\delta) + \xi^\alpha\eta^\beta(\sigma^\gamma\bar{\nu}_\gamma)(\tau^\delta\bar{\zeta}_\delta) \\
&\qquad + \phi^\alpha\psi^\beta(\lambda^\gamma\bar{\nu}_\gamma)(\kappa^\delta\bar{\zeta}_\delta)
\end{aligned} \tag{86}$$

Note, furthermore, that

$$\|\xi^{\alpha\cdots\beta\gamma\cdots\delta}\bar{\eta}_{\gamma\ldots\delta}\| \leq \|\xi^{\alpha\cdots\beta\gamma\cdots\delta}\|\|\bar{\eta}_{\gamma\ldots\delta}\|, \tag{87}$$

a result which is easily checked by placing $\xi^{\alpha\cdots\beta\gamma\cdots\delta}$ and $\bar{\eta}_{\gamma\ldots\delta}$ in normal form. Can this operation be extended from finite $\xi^{\alpha\cdots\beta\gamma\cdots\delta}, \bar{\eta}_{\gamma\ldots\delta}$ to the entire tensor product? To see that it can, let $\xi_i^{\alpha\cdots\beta\gamma\cdots\delta}$ and $\bar{\eta}_{i\gamma\ldots\delta}$, be Cauchy sequences of finite elements of $H^\alpha \otimes \ldots \otimes H^\beta \otimes H^\gamma \otimes \ldots \otimes H^\gamma$ and $\bar{H}_\gamma \otimes \ldots \otimes \bar{H}_\delta$, respectively. Then

$$\begin{aligned}
&\|\xi_i^{\alpha\cdots\delta}\bar{\eta}_{i\gamma\cdots\delta} - \xi_j^{\alpha\cdots\delta}\bar{\eta}_{j\gamma\cdots\delta}\| \\
&\quad = \|\xi_i^{\alpha\cdots\delta}\bar{\eta}_{i\gamma\cdots\delta} - \xi_j^{\alpha\cdots\delta}\bar{\eta}_{i\gamma\cdots\delta} + \xi_j^{\alpha\cdots\delta}\bar{\eta}_{i\gamma\cdots\delta} - \xi_j^{\alpha\cdots\delta}\bar{\eta}_{j\gamma\cdots\delta}\| \\
&\quad \leq \|\xi_i^{\alpha\cdots\delta} - \xi_j^{\alpha\cdots\delta}\|\|\bar{\eta}_{i\gamma\cdots\delta}\| + \|\xi_j^{\alpha\cdots\delta}\|\|\bar{\eta}_{i\gamma\cdots\delta} - \bar{\eta}_{j\gamma\cdots\delta}\|
\end{aligned} \tag{88}$$

where we have used (87). Thus, $\xi^{\alpha\cdots\beta\gamma\cdots\delta}\bar{\eta}_{\gamma\ldots\delta}$ is a Cauchy sequence in the Hilbert space $H^\alpha \otimes \ldots \otimes H^\beta$; hence it converges to some element of this Hilbert space. In this way, we associate with any (not necessarily finite) elements

$\xi^{\alpha\cdots\beta\gamma\cdots\delta}$ and $\bar{\eta}_{\gamma\cdots\delta}$ an element $\xi^{\alpha\cdots\beta\gamma\cdots\delta}\bar{\eta}_{\gamma\cdots\delta}$. This product is clearly linear in the factors, and satisfies (87). The operation of contraction is thus extended from the Hilbert spaces to their tensor products. Note that (85) is now a special case.

Now let $\xi^{\alpha\cdots\beta}$ and $\eta^{\gamma\cdots\delta}$ be finite elements of $H^\alpha\otimes\ldots\otimes H^\beta$ and $H^\gamma\otimes\ldots\otimes H^\delta$, respectively. We can certainly associate with these two an element, $\xi^{\alpha\cdots\beta}\eta^{\gamma\cdots\delta}$, $H^\alpha\otimes\ldots\otimes H^\beta\otimes H^\gamma\otimes\ldots\otimes H^\delta$. For example,

$$(\xi^\alpha\eta^\beta+\sigma^\alpha\tau^\beta)(\phi^\gamma\psi^\delta+\lambda^\gamma\kappa^\delta)=\xi^\alpha\eta^\beta\phi^\gamma\psi^\delta+\sigma^\alpha\tau^\beta\phi^\gamma\psi^\delta+\xi^\alpha\eta^\beta\lambda^\gamma\kappa^\delta+\sigma^\alpha\tau^\beta\lambda^\gamma\kappa^\delta \tag{89}$$

Note, furthermore, that

$$\|\xi^{\alpha\cdots\beta}\eta^{\gamma\cdots\delta}\|=\|\xi^{\alpha\cdots\beta}\|\|\eta^{\gamma\cdots\delta}\| \tag{90}$$

This operation, too, can be extended from finite elements to the entire tensor product. Let $\xi_i^{\alpha\cdots\beta}$ and $\eta_i^{\gamma\cdots\delta}$ be Cauchy sequences. Then $\xi_i^{\alpha\cdots\beta}\eta_i^{\gamma\cdots\delta}$ is also a Cauchy sequence, for

$$\begin{aligned}&\|\xi_i^{\alpha\cdots\beta}\eta_i^{\gamma\cdots\delta}-\xi_j^{\alpha\cdots\beta}\eta_j^{\gamma\cdots\delta}\|\\&\quad=\|\xi_i^{\alpha\cdots\beta}\eta_i^{\gamma\cdots\delta}-\xi_j^{\alpha\cdots\beta}\eta_i^{\gamma\cdots\delta}+\xi_j^{\alpha\cdots\beta}\eta_i^{\gamma\cdots\delta}-\xi_j^{\alpha\cdots\beta}\eta_j^{\gamma\cdots\delta}\|\\&\quad\le\|\xi_i^{\alpha\cdots\beta}-\xi_j^{\alpha\cdots\beta}\|\|\eta_i^{\gamma\cdots\delta}\|+\|\xi_j^{\alpha\cdots\beta}\|\|\eta_i^{\gamma\cdots\delta}-\eta_j^{\gamma\cdots\delta}\|\end{aligned} \tag{91}$$

This Cauchy sequence must converge to some element of $H^\alpha\otimes\ldots\otimes H^\beta\otimes H^\gamma\otimes\ldots\otimes H^\delta$. Thus, with any (not necessarily finite) elements $\xi^{\alpha\cdots\beta}$ and $\eta^{\gamma\cdots\delta}$ we may associate an element $\xi^{\alpha\cdots\beta}\eta^{\gamma\cdots\delta}$. This product is linear in the factors and satisfies (90). The operation of "outer product" is thus defined on our tensor product spaces. Note that the "formal products" which appear in (71) now have operational significance. (This informalization is typical of the final status of formal operations.)

We remark that our discussion of outer products above is merely the "finite part" of a more general result:

$$(H^\alpha\otimes\ldots\otimes H^\beta)\otimes(H^\gamma\otimes\ldots\otimes H^\delta)=H^\alpha\otimes\ldots\otimes H^\beta\otimes H^\gamma\otimes\ldots\otimes H^\delta \tag{92}$$

That is, we have shown that finite elements of the left side can be considered as elements of the right side of (92). The rest of the proof is analogous to the proofs above.

We next consider the extension of operators from our Hilbert spaces to their tensor products. Let A be a linear operator from H^α to H^γ. This operator can be written $A^\gamma{}_\alpha$: the result of acting on $\xi^\alpha\in H^\alpha$ with A is written $A^\gamma{}_\alpha\xi^\alpha$. (The notation may be misleading here. While every element of $H^\gamma\otimes\bar{H}_\alpha$ defines such an operator, not every operator can be considered as belonging to $H^\gamma\otimes\bar{H}_\alpha$.) With each finite element $\xi^{\alpha\cdots\beta}$ of $H^\alpha\otimes\ldots\otimes H^\beta$ we may certainly associate an element, $A^\gamma{}_\alpha\xi^{\alpha\cdots\beta}$, of $H^\gamma\otimes\ldots\otimes H^\beta$. For example,

$$A^\gamma{}_\alpha(\phi^\alpha\psi^\beta+\lambda^\alpha\kappa^\beta)=(A^\gamma{}_\alpha\phi^\alpha)\psi^\beta+(A^\gamma{}_\alpha\lambda^\alpha)\kappa^\beta \tag{93}$$

Unfortunately, there is in general no inequality which will permit us to extend this operation to the entire Hilbert space $H^\alpha \otimes \ldots \otimes H^\beta$. Thus, in general, $A^\gamma{}_\alpha \xi^{\alpha\cdots\beta}$ will only be defined for finite $\xi^{\alpha\cdots\beta}$; these elements, fortunately, are dense in $H^\alpha \otimes \ldots \otimes H^\beta$. There is, however, one condition under which the operation can be extended to the entire Hilbert space $H^\alpha \otimes \ldots \otimes H^\beta$. If $A^\gamma{}_\alpha$ is bounded, then

$$\|A^\gamma{}_\alpha \xi^{\alpha\cdots\beta}\| \leq |A| \|\xi^{\alpha\cdots\beta}\|. \tag{94}$$

(Proof: Use normal form.) In this case, (94) implies that $A^\gamma{}_\alpha \xi_i^{\alpha\cdots\beta}$ is a Cauchy sequence if $\xi_i^{\alpha\cdots\beta}$ is, and so we may define $A^\gamma{}_\alpha \xi^{\alpha\cdots\beta}$ for any $\xi^{\alpha\cdots\beta}$.

In fact, in many applications of the tensor product, the H^α, H^β, etc. are all merely copies of one fixed Hilbert space H. We then have one-to-one correspondences between H^α, H^β, etc. These correspondences can be indicated by retaining the root letter. For example, the element of H^β corresponding with $\xi^\alpha \in H^\alpha$ would be written ξ^β. The correspondences between our underlying Hilbert spaces induce correspondences, in an obvious way, between various tensor products. We may thus give meaning to such expressions as $\xi^{\alpha\beta} + \xi^{\beta\alpha}$. In this case — when all our underlying Hilbert spaces are copies of a Hilbert space H — we may introduce symmetrization over tensor indices (round brackets), e.g.,

$$\xi^{(\alpha\beta\gamma)} = \frac{1}{6}\left(\xi^{\alpha\beta\gamma} + \xi^{\beta\gamma\alpha} + \xi^{\gamma\alpha\beta} + \xi^{\beta\alpha\gamma} + \xi^{\gamma\beta\alpha} + \xi^{\alpha\gamma\beta}\right) \tag{95}$$

and anti-symmetrization over tensor indices (square brackets), e.g.,

$$\xi^{[\alpha\beta\gamma]} = \frac{1}{6}\left(\xi^{\alpha\beta\gamma} + \xi^{\beta\gamma\alpha} + \xi^{\gamma\alpha\beta} - \xi^{\beta\alpha\gamma} - \xi^{\gamma\beta\alpha} - \xi^{\alpha\gamma\beta}\right) \tag{96}$$

Note that any Cauchy sequence of symmetric (resp., skew) tensors of a given rank converges to a tensor which is necessarily symmetric (resp., skew). Hence, the symmetric (resp., skew) tensors of a given rank themselves form a Hilbert space. Similar remarks apply, of course, to any other symmetry on tensor indices.

There are an enormous number of facts about tensor products of Hilbert spaces. We have stated a few of them — and proven still fewer — here. It is the sheer bulk of the information, however, which makes the index notation valuable. Elementary facts are made to look elementary, and the mind is freed for important questions.

10 Fock Space: The Symmetric Case

The arena in which one discusses systems of many noninteracting identical particles is a Hilbert space called Fock space. This Hilbert space is constructed in terms of the Hilbert space H of one-particle states. Although the construction of H itself depends on the type of particle being considered (neutrinos, electrons, mesons, photons, etc.), the steps leading from H to its Fock space are independent of such details. In fact, there are two Fock spaces which can be associated

with a given Hilbert space H — what we shall call the symmetric Fock space and the anti-symmetric Fock space. If H represents the one-particle states of a Boson field, the appropriate space of many-particle states is the symmetric Fock space based on H. Similarly, fermions are described by the anti-symmetric Fock space. We shall define the Fock spaces associated with a Hilbert space H and a few of the operators on these spaces.

Let H be a Hilbert space. The *(symmetric) Fock space* based on H is the Hilbert space

$$\mathbb{C} \oplus H^{\alpha} \oplus (H^{(\alpha} \otimes H^{\beta)}) \oplus (H^{(\alpha} \otimes H^{\beta} \otimes H^{\gamma)}) \oplus \cdots \tag{97}$$

where H^{α}, H^{β}, etc. are all copies of H (Sect. 9), and where the round brackets surrounding the indices of the tensor products mean that the Hilbert space of symmetric tensors is to be used. More explicitly, an element of the symmetric Fock space consists of a string

$$\Psi = (\xi, \xi^{\alpha}, \xi^{\alpha\beta}, \xi^{\alpha\beta\gamma}, \ldots) \tag{98}$$

where ξ is a complex number, ξ^{α} is an element of H, $\xi^{\alpha\beta}$ is a symmetric ($\xi^{\alpha\beta} = \xi^{(\alpha\beta)}$) second-rank tensor over H, $\xi^{\alpha\beta\gamma}$ is a symmetric third-rank tensor over H, etc., for which the sum

$$\|\Psi\|^2 = \xi\bar{\xi} + \xi^{\alpha}\bar{\xi}_{\alpha} + \xi^{\alpha\beta}\bar{\xi}_{\alpha\beta} + \xi^{\alpha\beta\gamma}\bar{\xi}_{\alpha\beta\gamma} + \cdots , \tag{99}$$

which defines the norm of Ψ, converges. Physically, $\xi^{\alpha_1 \cdots \alpha_n}$ represents the "n-particle contribution" to Ψ. That the tensors are required to be symmetric is a reflection of the idea that "Ψ is invariant under interchange of identical particles".

We next introduce the creation and annihilation operators. Let $\sigma \in H$. We associate with this σ an operator $C(\sigma)$ on Fock space, this operator defined by its action on a typical element (98):

$$C(\sigma)\Psi = (0, \sigma^{\alpha}\xi, \sqrt{2}\sigma^{(\alpha}\xi^{\beta)}, \sqrt{3}\sigma^{(\alpha}\xi^{\beta\gamma)}, \ldots) \tag{100}$$

Similarly, with each $\bar{\tau} \in \bar{H}$ we associate an operator $A(\bar{\tau})$, defined by

$$A(\bar{\tau})\Psi = (\xi^{\mu}\bar{\tau}_{\mu}, \sqrt{2}\xi^{\mu\alpha}\bar{\tau}_{\mu}, \sqrt{3}\xi^{\mu\alpha\beta}\bar{\tau}_{\mu}, \ldots) \tag{101}$$

This $C(\sigma)$ is called the *creation operator* (associated with σ); $A(\bar{\tau})$ the *annihilation operator* (associated with $\bar{\tau}$). Note that the creation and annihilation operators are only defined on a dense subset of Fock space, for, in general, the sum on the right in (99) will not converge for the right sides of (100) and (101). It is an easy exercise in tensor calculus to work out the commutators of these operators:

$$\begin{aligned} [C(\sigma), C(\sigma')] &= 0 \\ [A(\bar{\tau}), A(\bar{\tau}')] &= 0 \\ [A(\bar{\tau}), C(\sigma)] &= (\sigma^{\mu}\bar{\tau}_{\mu})\mathbb{I} \end{aligned} \tag{102}$$

For example, the last equation in (102) would be derived as follows:

$$\begin{aligned} A(\bar\tau)C(\sigma)\Psi &= A(\bar\tau)(0,\sigma^\alpha\xi,\sqrt{2}\sigma^{(\alpha}\xi^{\beta)},\sqrt{3}\sigma^{(\alpha}\xi^{\beta\gamma)},\ldots) \\ &= (\xi\sigma^\mu\bar\tau_\mu,\xi^\alpha\sigma^\mu\bar\tau_\mu+\sigma^\alpha\xi^\mu\bar\tau_\mu,\xi^{\alpha\beta}\sigma^\mu\bar\tau_\mu+2\sigma^{(\alpha}\xi^{\beta)\mu}\bar\tau_\mu,\ldots) \\ C(\sigma)A(\bar\tau)\Psi &= C(\sigma)(\xi^\mu\bar\tau_\mu,\sqrt{2}\xi^{\mu\alpha}\bar\tau_\mu,\sqrt{3}\xi^{\mu\alpha\beta}\bar\tau_\mu,\ldots) \\ &= (0,\sigma^\alpha\xi^\mu\bar\tau_\mu,2\sigma^{(\alpha}\xi^{\beta)\mu}\bar\tau_\mu,3\sigma^{(\alpha}\xi^{\beta\gamma)\mu}\bar\tau_\mu,\ldots) \end{aligned} \tag{103}$$

Furthermore, if ϕ represents the element $(\eta,\eta^\alpha,\eta^{\alpha\beta},\ldots)$ of Fock space, we have

$$(C(\sigma)\Psi,\phi)=(\Psi,A(\bar\sigma)\phi) \tag{104}$$

for both sides of this equation are given by the sum

$$\xi\sigma^\mu\bar\eta_\mu+\sqrt{2}\xi^\alpha\sigma^\beta\bar\eta_{\alpha\beta}+\sqrt{3}\xi^{\alpha\beta}\sigma^\gamma\bar\eta_{\alpha\beta\gamma}+\cdots \tag{105}$$

Eqn. (104) is often summarized in words by saying that $C(\sigma)$ and $A(\bar\sigma)$ are *adjoints* of each other. (An operator is thus its own adjoint if and only if it is Hermitian. Technical distinctions are sometimes made between the adjoint and the Hermitian conjugate, and between self-adjoint and Hermitian. We shall not make these distinctions until they arise.) We can now understand why the strange factors $\sqrt{2}$, $\sqrt{3}$, etc. were inserted in the definitions (100) and (101). These are the only positive real factors for which the resulting creation and annihilation operators will satisfy (102) and (104).

We next introduce two number operators. For $\sigma\neq 0$, the *number operator in the state* σ is defined by

$$N(\sigma)=\|\sigma\|^{-2}C(\sigma)A(\bar\sigma) \tag{106}$$

whence

$$N(\sigma)\Psi=\|\sigma\|^{-2}(0,\sigma^\alpha\xi^\mu\bar\sigma_\mu,2\sigma^{(\alpha}\xi^{\beta)\mu}\bar\sigma_\mu,\ldots) \tag{107}$$

The *total number operator*, N, is defined by

$$N\Psi=(0,\xi^\alpha,2\xi^{\alpha\beta},3\xi^{\alpha\beta\gamma},4\xi^{\alpha\beta\gamma\delta},\ldots) \tag{108}$$

Note that these operators, too, are only defined on a dense subset of Fock space. We can think, intuitively, of N as resulting from "summing the $N(\sigma)$'s over an orthonormal basis", using

$$\text{“}\sum_{\substack{\text{orthonormal}\\ \text{basis}}}\sigma^\alpha\bar\sigma_\mu=\delta^\alpha{}_\mu\text{”} \tag{109}$$

where $\delta^\alpha{}_\mu$ is the "Kronecker delta". The number operators, $N(\sigma)$ and N, are Hermitian. Various commutators are as follows:

$$\begin{aligned} [N(\sigma),C(\sigma)]&=[N,C(\sigma)]=C(\sigma) \\ [N(\sigma),A(\bar\sigma)]&=[N,A(\bar\sigma)=-A(\bar\sigma) \\ [N(\sigma),N]&=0 \end{aligned} \tag{110}$$

(Commutators involving $N(\sigma)$ are most easily evaluated using (106) and (102), rather than (107), (100), and (101). Furthermore, one can guess the commutators involving N using the intuitive remark surrounding (109).)

Finally, we write down the eigenvectors and eigenvalues of our number operators. Clearly, from (108), the eigenvalues of N are precisely the non-negative integers: $0, 1, 2, \ldots$. The most general eigenvector with eigenvalue n is:

$$(0, 0, \ldots, 0, \xi^{\alpha_1 \cdots \alpha_n}, 0, \ldots) \tag{111}$$

The eigenvectors of $N(\sigma)$ are only slightly more difficult. First note that a tensor $\xi^{\alpha_1 \cdots \alpha_n}$ satisfies

$$\sigma^{(\alpha_1} \xi^{\alpha_2 \cdots \alpha_n) \mu} \bar{\sigma}_\mu = \nu \xi^{\alpha_1 \cdots \alpha_n} \tag{112}$$

for some complex number ν if and only if

$$\xi^{\alpha_1 \cdots \alpha_n} = \sigma^{(\alpha_1} \cdots \sigma^{\alpha_m} \kappa^{\alpha_{m+1} \cdots \alpha_n)} \tag{113}$$

for some $\kappa^{\alpha_{m+1} \cdots \alpha_n}$ satisfying $\kappa^{\alpha_{m+1} \cdots \alpha_n} \bar{\sigma}_{\alpha_n} = 0$. Proof: If $\nu = 0$, we're through. If $\nu \neq 0$, (112) implies

$$\xi^{\alpha_1 \cdots \alpha_n} = \sigma^{(\alpha_1} \lambda^{\alpha_2 \ldots \alpha_n)} \tag{114}$$

for some $\lambda^{\alpha_2 \ldots \alpha_n}$. If $\lambda^{\alpha_2 \ldots \alpha_n} \bar{\sigma}_{\alpha_n} = 0$, we're through. If not, substitute (114) into (112) to obtain

$$\xi^{\alpha_1 \cdots \alpha_n} = \sigma^{(\alpha_1} \sigma^{\alpha_2} \rho^{\alpha_3 \ldots \alpha_n)} \tag{115}$$

Continue in this way. It is now clear, from (107), that the most general simultaneous eigenvector of N and $N(\sigma)$, with eigenvalues n and m, respectively, is

$$(0, 0, \ldots, 0, \sigma^{(\alpha_1} \cdots \sigma^{\alpha_m} \kappa^{\alpha_{m+1} \cdots \alpha_n)}, 0, \ldots) \tag{116}$$

where $\kappa^{\alpha_{m+1} \cdots \alpha_n} \bar{\sigma}_{\alpha_n} = 0$.

11 Fock Space: The Anti-Symmetric Case

The definition and properties of Fock space in the antisymmetric case are closely analogous to those in the symmetric case.

Let H be a Hilbert space. The *(anti-symmetric) Fock space* based on H is the Hilbert space

$$\mathbb{C} \oplus H^\alpha \oplus (H^{[\alpha} \otimes H^{\beta]}) \oplus (H^{[\alpha} \otimes H^\beta \otimes H^{\gamma]}) \oplus \cdots \tag{117}$$

where H^α, H^β, etc. are all copies of H, and where the square brackets surrounding the indices of the tensor products mean that the Hilbert space of anti-symmetric tensors is to be used. That is, an element of the antisymmetric Fock space consists of a string

$$\Psi = (\xi, \xi^\alpha, \xi^{\alpha\beta}, \xi^{\alpha\beta\gamma}, \ldots) \tag{118}$$

of anti-symmetric tensors over H for which the sum

$$\|\Psi\|^2 = \xi\bar{\xi} + \xi^\alpha \bar{\xi}_\alpha + \xi^{\alpha\beta}\bar{\xi}_{\alpha\beta} + \cdots , \tag{119}$$

which defines the norm of Ψ, converges. That the tensors are required to he anti-symmetric is a reflection of the idea that "Ψ reverses sign under the interchange of identical particles." Physically, $\xi^{\alpha_1\cdots\alpha_n}$ represents the "n-particle contribution" to Ψ.

We associate with each $\sigma \in H$ a *creation operator*, $C(\sigma)$, and with each $\bar{\tau} \in \bar{H}$ an *annihilation operator*, $A(\bar{\tau})$, on Fock space as follows:

$$C(\sigma)\Psi = (0, \sigma^\alpha\xi, \sqrt{2}\sigma^{[\alpha}\xi^{\beta]}, \sqrt{3}\sigma^{[\alpha}\xi^{\beta\gamma]}, \ldots) \tag{120}$$

$$A(\bar{\tau})\Psi = (\xi^\mu\bar{\tau}_\mu, \sqrt{2}\xi^{\mu\alpha}\bar{\tau}_\mu, \sqrt{3}\xi^{\mu\alpha\beta}\bar{\tau}_\mu, \ldots) \tag{121}$$

As in the symmetric case, these operators are only defined on a dense subset of Fock space. The commutators of these creation and annihilation operators certainly exist — but they don't reduce to anything simple. We define the *anti-commutator* of two operators:

$$\{A, B\} = AB + BA \tag{122}$$

It is the anti-commutators of the creation and annihilation operators which are simple in the antisymmetric case:

$$\begin{aligned} \{C(\sigma), C(\sigma')\} &= 0 \\ \{A(\bar{\tau}), C(\bar{\tau}')\} &= 0 \\ \{A(\bar{\tau}), C(\sigma)\} &= (\sigma^\mu\bar{\tau}_\mu)\mathbb{I} \end{aligned} \tag{123}$$

The creation and annihilation operators are still adjoints of each other:

$$(C(\sigma)\Psi, \phi) = (\Psi, A(\bar{\sigma})\phi) \tag{124}$$

There is one further property of the creation and annihilation operators which is special to the antisymmetric case. Setting $\sigma = \sigma'$ in (123), we obtain:

$$C(\sigma)^2 = 0 \qquad\qquad A(\bar{\tau})^2 = 0 \tag{125}$$

These equations have a simple physical interpretation. If we try to create two particles in the same state, or annihilate two particles from the same state, we obtain zero. That is to say, one "can't have more than one particle in a given state." This, of course, is the essential feature of Fermi statistics.

The *number operator in the state* σ ($\neq 0$) and *total number operator* are defined by:

$$N(\sigma)\Psi = \|\sigma\|^{-2}C(\sigma)A(\bar{\sigma})\Psi = (0, \sigma^\alpha\xi^\mu\bar{\sigma}_\mu, 2\sigma^{[\alpha}\xi^{|\mu|\beta]}\bar{\sigma}_\mu, \ldots) \tag{126}$$

$$N\Psi = (0, \xi^\alpha, 2\xi^{\alpha\beta}, 3\xi^{\alpha\beta\gamma}, \ldots) \tag{127}$$

These operators are Hermitian. We can think of N as obtained by "summing the $N(\sigma)$'s over an orthonormal basis." Some commutators are:

$$\begin{aligned} [N(\sigma), C(\sigma)] &= [N, C(\sigma)] = C(\sigma) \\ [N(\sigma), A(\bar{\sigma})] &= [N, A(\bar{\sigma})] = -A(\bar{\sigma}) \\ [N(\sigma), N] &= 0 \end{aligned} \tag{128}$$

(It is interesting that one must use commutators, and not anticommutators, to make (128) simple.) The number operator in the state σ has one further property, this one special to the antisymmetric case. From (126), (123), and (125), we have

$$N(\sigma)^2 = N(\sigma) \tag{129}$$

Clearly, (129) is again saying that "occupation numbers in the antisymmetric case are either 0 or 1." A Hermitian operator which is equal to its square is called a *projection operator*. Eqn. (129) (or, alternatively, Eqn. (126)) implies that $N(\sigma)$ is bounded (and, in fact, $|N(\sigma)| = 1$). Hence, from Section 7, $N(\sigma)$ is defined on all of Fock space. On the other hand, N is only defined on a dense subset.

Finally, we write down the eigenvectors and eigenvalues of our number operators. Once again, the eigenvalues of N are the nonnegative integers, and the general eigenvector with eigenvalue n is:

$$(0, 0, \ldots, 0, \xi^{\alpha_1 \cdots \alpha_n}, 0, \ldots) \tag{130}$$

The eigenvalue-eigenvector structure of $N(\sigma)$, however, is quite different from that of the symmetric case. In fact, (129) implies that the only eigenvalues of $N(\sigma)$ are 0 and 1. First note that

$$\sigma^{[\alpha_1} \xi^{|\mu|\alpha_2 \cdots \alpha_n]} \bar{\sigma}_\mu = \nu \xi^{\alpha_1 \cdots \alpha_n} \tag{131}$$

if and only if either $\xi^{\alpha_1 \cdots \alpha_n} \bar{\sigma}_{\alpha_n}$ or

$$\xi^{\alpha_1 \cdots \alpha_n} = \sigma^{[\alpha_1} \kappa^{\alpha_2 \cdots \alpha_n]} \tag{132}$$

for some tensor $\kappa^{\alpha_2 \cdots \alpha_n}$ which is antisymmetric. (We needn't require, in addition, that $\kappa^{\alpha_2 \cdots \alpha_n} \bar{\sigma}_{\alpha_n} = 0$. Any multiples of σ^α which are contained in $\kappa^{\alpha_2 \cdots \alpha_n}$ will be lost in (132) because of the anti-symmetrization on the right.) Thus the most general eigenvector of $N(\sigma)$ with eigenvalue 0 is (118) with

$$\xi^\alpha \bar{\sigma}_\alpha = 0, \quad \xi^{\alpha\beta} \bar{\sigma}_\alpha, \quad \xi^{\alpha\beta\gamma} \bar{\sigma}_\alpha = 0, \quad \ldots \tag{133}$$

The most general eigenvector with eigenvalue 1 is

$$(0, \sigma^\alpha \kappa, \sigma[\alpha \kappa^{\beta]}, \sigma^{[\alpha} \kappa^{\beta\gamma]}, \sigma^{[\alpha} \kappa^{\beta\gamma\delta]}, \ldots). \tag{134}$$

12 Klein-Gordon Fields as Operators

Everybody knows that one essential idea of quantum field theory is that classical fields (e.g., real or complex-valued functions of position in Minkowski space) are to become operators (operator-valued functions of position in Minkowski space) on some Hilbert space. We have now assembled enough machinery to discuss this transition from fields to operators in the Klein-Gordon case. Of course, the same program will have to be carried out later — in essentially the same way — for other fields. The resulting field operators will play an important role when we discuss interactions.

Let ϕ^+ be a positive-frequency solution of the Klein-Gordon equation. That is, ϕ^+ is a complex-valued function of position in Minkowski space. The complex-conjugate function of ϕ^+, i.e., the function defined by the property that its value at a point in Minkowski space is to be the complex-conjugate of the value of ϕ^+, will be written ϕ^-. Clearly, ϕ^- is a negative-frequency solution of the Klein-Gordon equation. Finally, we introduce the real solution

$$\phi(x) = \phi^+(x) + \phi^-(x) \tag{135}$$

of the Klein-Gordon equation. The functions ϕ^+ and $\phi-$ can certainly be recovered from ϕ: they are the positive- and negative-frequency parts, respectively, of ϕ. Alternatively, these relations can be discussed in terms of functions in momentum space. Let h^+ (resp., h^-) be the complex-valued function on the future mass shell M_μ^+, (resp., the past mass shell M_μ^+) which represents the positive-frequency (resp., negative-frequency) solution ϕ^+ (resp., ϕ^-) of the Klein-Gordon equation. Then h^+ and h^- are clearly related as follows:

$$h^-(k) = \overline{h^+(-k)} \tag{136}$$

That is to say, the value of h^- at $k_a \in M_\mu^-$ is the complex-conjugate of the value of h^+ at $(-k_a) \in M_\mu^+$. Finally, the function h in momentum space which represents the real solution ϕ is given by

$$h = h^+ + h^- \tag{137}$$

This h has the property that it is invariant under simultaneous complex-conjugation and reflection through the origin, a property equivalent to the reality of ϕ.

Our ϕ^+, ϕ^-, and ϕ each assign a number (for the first two, a complex number; for the last, a real number) to each point of Minkowski space. Roughly speaking, what we want to do is find objects $\underline{\phi}^+(x)$, $\underline{\phi}^-(x)$, and $\underline{\phi}(x)$ which assign, to each point in Minkowski space, an operator. These operators are to act on the (symmetric) Fock space based on the Hilbert space of positive-frequency solutions of the Klein-Gordon equation. The relationship between the functions $\phi^+(x)$, $\phi^-(x)$, and $\phi(x)$ is to be reflected, in an appropriate way, as a relationship between the corresponding operators. Since the functions $\phi^+(x)$ and $\phi^-(x)$ are complex-conjugates of each other, we demand that the operators $\underline{\phi}^+(x)$ and

$\underline{\phi}^-(x)$ be adjoints of each other; since the function $\phi(x)$ is real and the sum of $\phi^+(x)$ and $\phi^-(x)$, we demand that the operator $\underline{\phi}(x)$ be Hermitian and the sum of the operators $\underline{\phi}^+(x)$ and $\underline{\phi}^-(x)$. Still speaking roughly, one might imagine proceeding as follows. Choose an orthonormal basis σ_i $(i = 1, 2, \ldots)$ for the Hilbert space H of positive-frequency solutions of the Klein-Gordon equation. Then for each i we have a positive-frequency solution $\phi_i^+(x)$, a negative-frequency solution $\phi_i^-(x)$ (the complex-conjugate function of $\phi_i^+(x)$), and a real solution $\phi_i(x)$ (the sum of $\phi_i^+(x)$ and $\phi_i^-(x)$). Then any triple of solutions, $\phi^+(x)$, $\phi^-(x)$, and $\phi(x)$, related as above (i.e., ϕ^+ is positive-frequency, ϕ^- is its complex-conjugate, and ϕ is their sum) could be expanded in terms of our basis:

$$\phi^+(x) = \sum_i a_i \phi_i^+(x) \tag{138}$$

$$\phi^-(x) = \sum_i \bar{a}_i \phi_i^-(x) \tag{139}$$

$$\phi(x) = \phi^+(x) + \phi^-(x) = \sum_i (a_i \phi_i^+(x) + \bar{a}_i \phi_i^-(x)) \tag{140}$$

Here, $a_1, a_2, \ldots$ are simply complex numbers. Thus, triples of solutions related as above would be characterized by sequences of complex numbers. To pass from fields to operators, we could now simply replace the coefficients in the expansions (138), (139), and (140) by the corresponding creation and annihilation operators:

$$\underline{\phi}^+(x) = \hbar \sum_i \phi_i^+(x) A(\sigma_i) \tag{141}$$

$$\underline{\phi}^-(x) = \hbar \sum_i \phi_i^-(x) C(\bar{\sigma}_i) \tag{142}$$

$$\underline{\phi}(x) = \underline{\phi}^+(x) + \underline{\phi}^-(x) = \hbar \sum (\phi_i^+(x) A(\sigma_i) + \phi_i^-(x) C(\bar{\sigma}_i)) \tag{143}$$

Fix the basis σ_i. Then, for each position x in Minkowski space, $\phi_i^+(x)$, $\phi_i^-(x)$, and $\phi_i(x)$ are just numbers. The right sides of (141), (142), and (143) are simply (infinite!) sums of operators on Fock space. In this way, we might expect to be able to associate operators, $\underline{\phi}^+(x)$, $\underline{\phi}^-(x)$, and $\underline{\phi}(x)$ with positions in Minkowski space. These operators would, of course, satisfy the appropriate adjoint, Hermiticity, and sum conditions. One further difference between (138), (139), (140) and (141), (142), (143) should be emphasized. Whereas (138), (139), (140) involve a *particular* solution of the Klein-Gordon equation, which is expanded in terms of a basis, no such particular solution is singled out in (141), (142), (143). The sums in (141), (142), (143) need be done just once to obtain operator-valued functions on Minkowski space.

The paragraph above merely outlines a program. In order to actually carry it out, we would, at least, have to solve three problems:

i) find a basis for our Hilbert space H whose elements can be represented as smooth (or at least continuous) functions on Minkowski space (recall

that an element of H is a function on the mass shell which need only be measurable and square-integrable),

ii) prove that the sums on the right of (141), (142), (143) converge in a suitable sense, and

iii) prove that the resulting operators are independent of the choice of basis.

The problem is with ii): in no reasonable sense do such sums converge. The way out of this difficulty is through the use of "smearing functions." The *support* of a function f on Minkowski space (more generally, on any topological space) is defined as the closure of the collection of points on which f is nonzero, so the support of f is a subset of Minkowski space. A smooth, real-valued function f with compact support will be called a *test function*. Note that, if G is any continuous function on Minkowski space, then the value of

$$G(f) = \int G(x) f(x)\, \mathrm{d}V \tag{144}$$

for every test function f determines uniquely the function G. (The integral (144) converges because f has compact support.) These remarks suggest that, instead of dealing with an operator-valued function $\phi(x)$ on Minkowski space — something we have not been able to obtain — we "smear out the x-dependence of the operators":

$$\underline{\phi}(f) = \int \underline{\phi}(x) f(x)\, \mathrm{d}V \tag{145}$$

That is to say, we consider our operators as depending, not on points in Minkowski space, but rather on the test functions themselves. Roughly speaking, $\underline{\phi}(f)$, for every test function f, has the same information content as $\underline{\phi}(x)$ for every point x of Minkowski space, just as $G(f)$ determines $G(x)$. The advantage of this formulation is that we may be able to define $\underline{\phi}(f)$, whereas we have not been able to define $\underline{\phi}(x)$. To obtain evidence on this question as to whether $\underline{\phi}(f)$ can be defined, we naively carry out the "smearing operation", (145), for (141), (142), (143):

$$\underline{\phi}^+(f) = \hbar \sum_i (\int \phi_i^+(x) f(x)\, \mathrm{d}V) A(\sigma_i) \tag{146}$$

$$\underline{\phi}^-(f) = \hbar \sum_i (\int \phi_i^-(x) f(x)\, \mathrm{d}V) C(\bar{\sigma}_i) \tag{147}$$

$$\underline{\phi}(f) = \underline{\phi}^+(f) + \underline{\phi}^-(f) \tag{148}$$

Thus, the operators which result from smearing are merely the creation and annihilation operators, "weighted by the components of f with respect to the basis vectors."

The preceding two paragraphs — particularly (146), (147), (148) — are intended to motivate the simple and precise definitions which follow. Let $f(x)$ be a test function. Then the Fourier inverse of f, $f'(k)$, is a smooth function

on all of momentum space (not just on the mass shell: $f(x)$ doesn't have to satisfy the Klein-Gordon equation). If, however, we restrict $f'(k)$ to the future mass shell, M_μ^+, we obtain a measurable and square-integrable function on M_μ^+, and hence an element, $\sigma(f)$, of our Hilbert space. (The proof that this $f'(k)$ is actually square-integrable over the mass shell is given in books on Fourier analysis.) That is, with each test function f there is associated an element $\sigma(f)$ of H. We now define:

$$\underline{\phi}^+(f) = \hbar A(\sigma(f)) \tag{149}$$

$$\underline{\phi}^-(f) = \hbar C(\overline{\sigma(f)}) \tag{150}$$

$$\underline{\phi}(f) = \underline{\phi}^+(f) + \underline{\phi}^-(f) = \hbar(A(\sigma(f)) + C(\overline{\sigma(f)})) \tag{151}$$

(See (146), (147), (148).) These are the smeared-out field operators. We remark that these field operators are linear mappings from the real vector space of test functions to the vector space of operators on Fock space. Note that $\underline{\phi}^+(f)$ and $\underline{\phi}^-(f)$ are adjoints of each other, and that $\underline{\phi}(f)$ is Hermitian with $\underline{\phi}(f) = \underline{\phi}^+(f) + \underline{\phi}^-(f)$. It is now clear why we were not able to define operators such as $\phi^+(x)$ earlier. We can think of $\underline{\phi}^+(x)$ as being the "limit of $\underline{\phi}^+(f)$ as f approaches a δ-function at x" (see (145).) But as f approaches a δ-function, $f'(k)$ approaches a function which is not square-integrable on the mass shell. Thus, $\underline{\phi}^+(x)$ represents "creation in an un-normalizable state."

We next wish to establish a result whose intuitive content is: "in their dependence on x, the operators $\underline{\phi}(x)$, $\underline{\phi}^-(x)$, and $\underline{\phi}(x)$ satisfy the Klein-Gordon equation, e.g.,

$$(\Box + \mu^2)\underline{\phi}^+(x) = 0." \tag{152}$$

Strictly speaking, of course, (152) has no meaning, for $\underline{\phi}^+(x)$ has no meaning. Is it possible, however, to express the essential content of (152) in terms of the well-defined operators (149)? Yes. The idea is to replace (152) with the statement that the result of smearing out the left side with an arbitrary test function f gives zero:

$$\int f(x)(\Box + \mu^2)\underline{\phi}^+(x)\,\mathrm{d}V = 0 \tag{153}$$

Formally integrating by parts twice, and throwing away surface terms since f has compact support, we are led to replace (153) (and hence (152)) by

$$\int \underline{\phi}^+(x)(\Box + \mu^2)f(x)\,\mathrm{d}V = 0 \tag{154}$$

These considerations suggest that we replace the meaningless equations (152) or (153) by an "equivalent" equation, (154), to which we can assign a meaning. We are thus led to the following conjecture: If f is a test function, then

$$\begin{aligned} \underline{\phi}^+((\Box + \mu^2)f) &= 0 \\ \underline{\phi}^-((\Box + \mu^2)f) &= 0 \\ \underline{\phi}((\Box + \mu^2)f) &= 0 \end{aligned} \tag{155}$$

(Note that if f is a test function, so is $(\Box + \mu^2)f$.) In fact, our conjecture is true, for if $f'(k)$ is the Fourier inverse of $f(x)$, then $(-k_a k^a + \mu^2)f'(k)$ is the Fourier inverse of $(\Box + \mu^2)f$. But $(-k_a k^a + \mu^2)f'(k)$ vanishes on the mass shell, so $\sigma((\Box + \mu^2)f) = 0$. Eqns. (155) now follow from (149), (150), (151). We conclude that, in an appropriate sense, our operator fields satisfy the Klein-Gordon equation.

Finally, we consider the commutators of our operators. Let f and g be test functions. Since any two creation operators commute with each other, and any two annihilation operators commute with each other (see (102)), we clearly have

$$[\underline{\phi}^+(f), \underline{\phi}^+(g)] = [\underline{\phi}^-(f), \underline{\phi}^-(g)] = 0 \tag{156}$$

Furthermore, since the commutator of any creation operator with any annihilation operator is a multiple of the identity operator (see (102)), we have

$$[\underline{\phi}^+(f), \underline{\phi}^-(g)] = \frac{\hbar}{i} D^+(f,g)\mathbb{I} \qquad [\underline{\phi}^-(f), \underline{\phi}^+(g)] = \frac{\hbar}{i} D^-(f,g)\mathbb{I} \tag{157}$$

where $D^+(f,g)$ and $D^-(f,g)$ are complex-valued (*not* operator-valued) functions of the test functions f and g. In fact, it follows immediately from (156), (149), (150), and (102) that

$$\begin{aligned} D^+(f,g) &= i\hbar\sigma^\alpha(g)\bar{\sigma}_\alpha(f) \\ D^-(f,g) &= -i\hbar\sigma^\alpha(f)\bar{\sigma}_\alpha(f) \end{aligned} \tag{158}$$

Therefore,

$$D^+(f,g) = -D^-(g,f) \tag{159}$$

$$D^+(f,g) = \overline{D^-(f,g)} \tag{160}$$

The commutators of the $\underline{\phi}$ operators follow from (151) and (157),

$$[\underline{\phi}(f), \underline{\phi}(g)] = \frac{\hbar}{i}(D^+(f,g) + D^-(f,g))\mathbb{I} = \frac{\hbar}{i} D(f,g)\mathbb{I} \tag{161}$$

where the second equality is the definition of $D(f,g)$. Eqns. (159) and (160) now imply

$$D(f,g) = -D(g,f) \tag{162}$$

$$D(f,g) = \overline{D(f,g)} \tag{163}$$

The D-functions have one further property, which can be called *Poincaré invariance*. Let $x \to Px$ be a Poincaré transformation on Minkowski space which does not reverse the future and past time directions. (This last stipulation is necessary because the distinction between positive frequency and negative frequency requires a particular choice of a "future" time direction on Minkowski space.) Then, defining the test functions $\tilde{f}(x) = f(Px)$, $\tilde{g}(x) = g(Px)$, we have

$$D^\pm(\tilde{f}, \tilde{g}) = D^\pm(f,g) \qquad D(\tilde{f}, \tilde{g}) = D(f,g) \tag{164}$$

The functions $D^{\pm}(f,g)$ and $D(f,g)$ are often discussed by introducing distributions $D^{\pm}(x,y)$ and $D(x,y)$ on Minkowski space, and setting

$$\begin{aligned} D^{\pm}(f,g) &= \int \mathrm{d}V_x \int \mathrm{d}V_y \, D^{\pm}(x,y) f(x) g(y) \\ D(f,g) &= \int \mathrm{d}V_x \int \mathrm{d}V_y \, D(x,y) f(x) g(y) \end{aligned} \tag{165}$$

It is not surprising that there should exist such distributions. A distribution, after all, is just a continuous linear mapping from the topological vector space of test functions to the reals, and $D^{\pm}(x,y)$ and $D(f,g)$ are certainly linear in their arguments. We shall not at this time discuss the topology on the space of test functions, nor prove that $D^{\pm}$ and D are continuous. Poincare invariance implies that

$$\begin{aligned} D^{\pm}(x,y) &= D^{\pm}(x-y) \\ D(x,y) &= D(x-y) \end{aligned} \tag{166}$$

where we have written $x-y$ for the position vector of x relative to y. It is not difficult to evaluate the functions (166) explicitly using (158) and a table of integrals. They involve Bessel functions.

There is, however, one particularly interesting property of $D(f,g)$. Test functions f and g will be said to have *relatively spacelike supports* if, for any point x of the support of f and any point y of the support of g, $x-y$ is spacelike. The property is the following: If f and g have relatively spacelike supports, then $D(f,g)=0$. The easiest proof is by means of the distribution $D(x,y)$. Eqn. (162) implies

$$D(x-y) = -D(y-x) \tag{167}$$

But if $x-y$ is spacelike, there is a Poincaré transformation which does not reverse future and past and which takes x to y and y to x (i.e., $x=Py$, $y=Px$). Poincaré invariance, (164), now implies

$$D(x-y) = D(y-x) \tag{168}$$

whence $D(x,y)=0$ for $x-y$ spacelike. That $D(f,g)=0$ when f and g have relatively spacelike supports now follows from (165).

13 The Hilbert Space of Solutions of Maxwell's Equations

We now wish to write down the quantum theory for a system of many free (non-interacting) photons. Our starting point is the classical field equations: Maxwell's equations. The method is entirely analogous to that of the Klein-Gordon equation: the electromagnetic field plays the role of the Klein-Gordon

field ϕ, the Maxwell equations the role of the Klein-Gordon equation. There are, of course, important differences between the two cases: a tensor field rather than a scalar field, two first-order tensor equations rather than one second-order scalar equation, etc. One further difference should be emphasized. Whereas the Klein-Gordon equation is, in a sense, the Schrödinger equation for a free particle, the Maxwell equations are classical (non-quantum). The electromagnetic analogy of the classical free particle, on the other hand, would be geometrical optics. Thus, we have the following table:

Electrodynamics	**Free Relativistic Particle**
Geometrical Optics	Classical Dynamics
Maxwell's Equations	Klein-Gordon Equation
Quantum Electrodynamics	Many-Particle Theory

(169)

Theories appearing in the same row are described, mathematically, in roughly the same terms: for the first row, curves in Minkowski space; for the second row, fields in Minkowski space; for the third row, creation and annihilation operators on Fock space.

The first step is to impose the structure of a Hilbert space on a certain collection of solutions of Maxwell's equations — just as we began the second-quantization of the Klein-Gordon equation by making a Hilbert space of solutions of that equation.

The *electromagnetic field* is a skew, second-rank tensor field F_{ab} ($= F_{[ab]}$) on Minkowski space. In the absence of sources, this field must satisfy *Maxwell's equations*:

$$\nabla_{[a}F_{bc]} = 0 \tag{170}$$

$$\nabla^a F_{ab} = 0 \tag{171}$$

Eqn. (170) implies that there exists a vector field A_a on Minkowski space for which

$$F_{ab}2\nabla_{[a}A_{b]} \tag{172}$$

Conversely, given any vector field A_a, the F_{ab} given by (172) satisfies (170). This A_a is called a *vector potential*. Substituting (172) into (171), we obtain

$$\Box A_a - \nabla_a(\nabla_b A^b) = 0 \tag{173}$$

Thus, any vector field satisfying (173) defines, via (172), a solution of Maxwell's equations, end, conversely, every solution of Maxwell's equations can be obtained from some vector potential satisfying (173). Two vector potentials, A_a and $\tilde{A}_a$, define (via (172)) the same F_{ab} if and only if

$$\tilde{A}_a = A_a + \nabla_a\Lambda \tag{174}$$

for some scalar field Λ on Minkowski space. Changes in the vector potential of the form (174) are called *gauge transformations*. By means of a gauge transformation one can find, for any solution of Maxwell's equations, a vector potential which satisfies

$$\nabla_a A^a = 0 \tag{175}$$

Vector potentials which satisfy (175) are said to be in the *Lorentz gauge*. If a vector potential for a solution of Maxwell's equations is in the Lorentz gauge, then, from (173), it satisfies

$$\Box A_a = 0 \tag{176}$$

If two vector potentials are both in the Lorentz gauge, and differ by a gauge transformation (174), then necessarily

$$\Box \Lambda = 0 \tag{177}$$

We can summarize the situation with the following awkward remark: the vector space of solutions of Maxwell's equations is equal to the quotient space of the vector space of vector fields which satisfy (175) and (176) by the vector subspace consisting of gradients of scalar fields which satisfy (177). All fields above are, of course, real.

We now do with Maxwell's equations what was done earlier with the Klein-Gordon equation: we go to momentum space. Let $A_a(x)$ be a vector potential, in the Lorentz gauge, for a solution to Maxwell's equations. Set

$$A_a(x) = \int\limits_M A'_a(k) e^{ik_b x^b} \, \mathrm{d}V \tag{178}$$

In (178), k represents position in momentum space, and $A'_a(k)$ associates a complex vector in momentum space with each such k. The integral on the right in (178) associates, with each point x in Minkowski space, a vector in momentum space, and hence a vector in Minkowski space at the point x. Thus, the right side of (178) defines a vector field in Minkowski space. We now demand that $A_a(x)$ given by (178) satisfy (176) and (175):

$$\Box A_a(x) = \int\limits_M (-k_c k^c) A'_a(k) e^{ik_b x^b} \, \mathrm{d}V = 0 \tag{179}$$

$$\nabla_a A^a(x) = \int\limits_M ik^a A'_a(k) e^{ik_b x^b} \, \mathrm{d}V = 0 \tag{180}$$

Eqn. (179) states that $A'_a(k)$ vanishes unless $k^a k_a = 0$. That is to say, $A'_a(k)$ need only be specified on the null cone in momentum space, or, what is the same thing, on the mass-zero shell, M_0. Thus, we can replace (178) by

$$A_a(x) = \int\limits_{M_0} A'_a(k) e^{ik_b x^b} \, \mathrm{d}V_0 \tag{181}$$

Eqn. (180) states that

$$k^a A'_a(k) = 0 \tag{182}$$

for every $k \in M_0$. An $A'_a(k)$ which satisfies (182) will be said to be *transverse*. Finally, the condition that $A(x)$, given by (181), be real, is

$$A'_a(-k) = \overline{A'_a(k)} \tag{183}$$

Eqn. (183) implies, in particular, that the knowledge of $A'_a(k)$ on M_0^+ determines uniquely the values of $A'_a(k)$ on M_0^-. We thus need only concern ourselves with $A'_a(k)$ on M_0^+.

To summarize, there is a one-to-one correspondence (modulo questions of convergence of Fourier integrals) between real vector fields $A_a(x)$ on Minkowski space which satisfy (175) and (176) and transverse complex vector functions $A'_a(k)$ on M_0^+.

Unfortunately, real vector fields $A_a(x)$ on Minkowski space which satisfy (175) and (176) are not the same as solutions of Maxwell's equations: we have to deal with the problem of gauge transformations. Let Λ satisfy (177), and let $\tilde{A}_a(x)$ be given by (174). Then the corresponding Fourier inverses, $\Lambda'(k)$ and $\tilde{A}'_a(k)$, are clearly related by:

$$\tilde{A}'_a(k) = A'_a(k) + ik_a\Lambda'(k) \tag{184}$$

In other words, a gauge transformation on $A_a(x)$ which preserves the Lorentz gauge corresponds simply to adding to $A'_a(k)$ a complex multiple of k_a. Note that, since k_a is null, the gauge transformations (184) do not destroy the transversality condition, (182).

To summarize, *there is a one-to-one correspondence* (modulo convergence of Fourier integrals) *between solutions of Maxwell's equations and equivalence classes of transverse complex vector functions $A'_a(k)$ on M_0^+, where two such functions $A'_a(k)$ are regarded as equivalent if they differ by a multiple of k_a.*

The reason for expressing the content of Maxwell's equations in terms of momentum space is that certain properties of the space of solutions of Maxwell's equations become more transparent there. We first impose on the (real!) solutions of Maxwell's equations the structure of a complex vector space. To add two solutions of Maxwell's equations, one simply adds the tensor fields on Minkowski space. Expressed in terms of momentum space, this means that one adds the corresponding $A'_a(k)$. To "multiply" a solution F_{ab} of Maxwell's equations by a complex number α, one multiplies the corresponding complex vector function $A'_a(k)$ by α in the usual way, and interprets the result, $\alpha A'_a(k)$, as a solution of Maxwell's equations (necessarily, a real solution). These operations clearly extend to operations on the equivalence classes of $A'_a(k)$, and hence are well-defined operations on solutions of Maxwell's equations. It is only when α is real that multiplying a solution F_{ab} by α, in the sense described above, is equivalent to simply multiplying the tensor field F_{ab} by α in the usual way. This cannot be the case, of course, when α is complex, for the usual product, αF_{ab}, would be a complex field on Minkowski space rather than a real one, and solutions of Maxwell's equation must he real. We can, however, give a picture for what the product of i and F_{ab} ("product" and "multiply" will always refer to that operation defined above) means. Let $A_a(x)$ be a vector potential in the Lorentz space, and let $A'_a(k)$ be as in (181). Then $iA'_a(k)$ corresponds to the

vector potential

$$\begin{aligned}\int\limits_{M_0^+} iA'_a(k)e^{ik_b x^b}\,\mathrm{d}V_0 - \int\limits_{M_0^-} i\,\overline{A'_a(k)}e^{ik_b x^b}\,\mathrm{d}V_0 \\ = \int\limits_{M_0^+} A'_a(k)e^{ik_b x^b+\frac{\pi}{2}}\,\mathrm{d}V_0 + \int\limits_{M_0^-} \overline{A'_a(k)}e^{ik_b x^b-\frac{\pi}{2}}\,\mathrm{d}V_0\end{aligned} \tag{185}$$

In other words, multiplication of a solution of Maxwell's equations by i corresponds to resolving F_{ab} into complex plane-waves, and shifting the phase of the positive frequency parts by $\pi/2$ while shifting the phase of the negative-frequency parts by $-\pi/2$. (In exactly the same way, the real solutions of the Klein-Gordon equation form a complex vector space.)

We next introduce an inner product on our complex vector space. We define the norm of a transverse complex vector function $A'_a(k)$ on M_0^+ by

$$\frac{2}{\hbar}\int\limits_{M_0^+} (-A'_a(k)\overline{A'^a(k)})\,\mathrm{d}V_0 \tag{186}$$

Since $A'_a(k)$ is transverse, and since $k^a k_a = 0$ on M_0, the real number (186) is clearly invariant under gauge transformations, (184), on $A'_a(k)$. Thus, the norm (186) is well-defined on solutions of Maxwell's equations. Furthermore, the norm (186) is non-negative and vanishes when and only when $A'_a(k) = 0$ (more properly, when and only when $A'_a(k)$ is in the zero equivalence class, i.e., when and only when $A'_a(k)$ is a multiple of k_a). To prove this, we show that the integrand is non-negative. Fix k, and let

$$A'_a(k) = m_a + in_a \tag{187}$$

where m_a and n_a are real. By transversality,

$$m_a k^a = n_a k^a = 0 \tag{188}$$

The integrand of (186) is

$$(-A'_a(k)\overline{A'^a(k)}) = -m^a m_a - n^a n_a \tag{189}$$

But (188) implies that m_a and n_a are either spacelike or multiples of k_a, whence (189) is nonnegative and vanishes when and only when $m_a + in_a$ is a multiple of k_a.

Thus, the collection of all equivalence classes of (say, continuous) transverse $A'_a(k)$ on for which (186) converges has the structure of an inner-product space. Its completion is our Hilbert space, H_M, of solutions of Maxwell's equations. Just as in the Klein-Gordon case, one can describe H_M directly in momentum space. It is the collection of all equivalence classes of measurable, square-integrable (in the sense that (186) converges), transverse $A'_a(k)$ on M_0^+.

This H_M represents the one-photon states. (Intuitively, a solution of Maxwell's equations represents a "wave function" for a single photon.) The space of many-photon states is the (symmetric, since photons are bosons) Fock space based on H_M. Thus, from our earlier discussion, we have creation, annihilation, and number operators for (free) photons. The commutation relations and other properties of these operators have already been worked out.

Finally, we introduce momentum operators on H_M. Let p^a be a constant vector field in Minkowski space. Then, with each solution $F_{ab}(x)$ of Maxwell's equations, we associate another solution: multiply the solution $-\hbar p^c \nabla_c F_{ab}$ of Maxwell's equations by the number i ("multiply", of course, in the sense of H_M). We thus define a linear operator, $P(p^a)$, on H^M. In momentum space, this operator clearly takes the form

$$P(p^b)A'_a(k) = \hbar(p^b k_b)A'_a(k) \tag{190}$$

Note that the momentum operators are only defined on a dense subset of H_M, are Hermitian, and commute with each other. Another interesting property of these operators — which also holds in the Klein-Gordon case — is that "energies are positive." Let p^a be timelike and future-directed, so $P(p^a)$ represents an energy operator. Then $p^a k_a \geq 0$ for any $k_a \in M_0^+$. Hence, from (190) and (186), the inner product of σ and $P(p^a)\sigma$ is positive for any element σ ($\neq 0$) of H_M.

Although they are not commonly discussed, one can also introduce position operators on H_M. As in the Klein-Gordon case, one projects to obtain a vector field on the mass shell. Instead of taking the directional derivative of a function on the mass shell as in the Klein-Gordon case, one takes the Lie derivative of $A'^a(k)$, considered as a contravariant vector field on M_0^+. (It's important, in order to preserve transversality, that one takes $A'^a(k)$ to be a contravariant rather than a covariant field.) Finally, one includes an appropriate "divergence-type term" in the operators in order to make them be Hermitian.

14 Maxwell Fields as Operators

We shall now introduce, on the Fock space for the Maxwell equation, operators associated with the classical fields, A_a and F_{ab}, of the Maxwell theory. The definitions are closely analogous to those of the Klein-Gordon theory.

Since the classical Klein-Gordon field ϕ is a scalar field, the test functions used to "smear out" the corresponding field operators are scalar fields. In the Maxwell case, on the other hand, the classical fields are vector or tensor fields on Minkowski space. One must therefore introduce "test functions" which themselves have vectorial or tensorial character. The *support* of a tensor field $f^{a_1 \cdots a_n}$ on Minkowski space is defined as the closure of the collection of all points of Minkowski Space at which $f^{a_1 \cdots a_n} \neq 0$. A smooth, real tensor field on Minkowski space, with compact support, will be called a *test field*. In order to facilitate calculations with such test fields, it is convenient to establish the following remark:

Lemma 1. *Let $T_{a_1\cdots a_n}$ be a smooth, totally antisymmetric tensor field on Minkowski space. Then*

$$\int T_{a_1\cdots a_n}\nabla^{[a_1}f^{a_2\cdots a_n]}\,\mathrm{d}V = 0 \tag{191}$$

for every totally antisymmetric test field $f^{a_2\cdots a_n}$ if and only if

$$\nabla^{a_1}T_{a_1a_2\cdots a_n} = 0 \tag{192}$$

Furthermore,

$$\int T_{a_1\cdots a_n}\nabla_m f^{ma_1\cdots a_n}\,\mathrm{d}V = 0 \tag{193}$$

for every totally antisymmetric test field $f^{ma_1\cdots a_n}$ if and only if

$$\nabla_{[m}T_{a_1\cdots a_n]} = 0 \tag{194}$$

Proof. Integrating by parts once, and discarding the surface term by compact support, we have the identity

$$\int T_{a_1\cdots a_n}\nabla^{[a_1}f^{a_2\cdots a_n]}\,\mathrm{d}V = -\int(\nabla^{a_1}T_{a_1\cdots a_n})f^{a_2\cdots a_n}\,\mathrm{d}V \tag{195}$$

for every totally antisymmetric test field $f^{a2\cdots a_n}$. But clearly the right side of (195) vanishes for every test field if and only if (192) holds. The second part of the Lemma is proved in the same way, using the identity

$$\int T_{a_1\cdots a_n}\nabla_m f^{ma_1\cdots a_n}\,\mathrm{d}V = -\int(\nabla_{[m}T_{a_1\cdots a_n]})f^{ma_1\cdots a_n}\,\mathrm{d}V \tag{196}$$

□

Note that Lemma 1 is easily generalized to higher order equations, to other symmetries of the tensors, etc. The essential idea is that linear differential equations on a tensor field $T_{a_1\cdots a_n}$ on Minkowski space can be expressed by the condition that the smeared-out version of this field vanish for an appropriate collection of test fields.

We begin with the field operators for the vector potential. Unfortunately, the classical vector potential, $A_a(x)$, is not determined uniquely by a solution of Maxwell's equations; there is the freedom of gauge transformations (174), where Λ is a solution of the wave equation. We would expect this gauge freedom to appear in some way in the corresponding operators. The essential observation is that, by Lemma 1, the quantity

$$\int A_a f^a\,\mathrm{d}V \tag{197}$$

is invariant under gauge transformations provided the test field f^a is the sum of a gradient and a vector field whose divergence vanishes. Conversely, the value of the real number (197) for every test field which is the sum of a gradient and

a divergence-free field determines $A_a(x)$ uniquely up to gauge transformations. We are thus led to view the gauge freedom in the vector potential as representing a restriction on the class of test fields which are appropriate for smearing out the vector potential.

The remarks above motivate the definitions below. Let f^a be a test field, and let $f'^a(k)$ be its Fourier inverse, a vector function on momentum space. Evidently, if f^a is divergence-free then

$$f'^a(k)k_a = 0 \tag{198}$$

while if f^a is a gradient then

$$f'^a(k) = h(k)k^a \tag{199}$$

It is clear, therefore, that if f^a is the sum of a gradient and a divergence-free field, then $f'^a(k)$, restricted to M_0^+, is transverse. In other words, we may associate, with each test field f^a on Minkowski space which is the sum of a gradient and a divergence-free field, an element $\sigma(f^a)$ of H_M. We define the vector potential operators

$$\underline{A}(f^a) = \hbar(C(\sigma(f^a)) + A(\overline{\sigma(f^a)}) \tag{200}$$

Note that the operator (200) is Hermitian, a result one expects because the corresponding classical field is real. The definition of the electromagnetic field operators is suggested by Lemma 1 and Eqn. (172). If f^{ab} is a skew test field, we define

$$\underline{F}(f^{ab}) = \underline{A}(2\nabla_b f^{ab}) \tag{201}$$

Thus, the electromagnetic field operators (which are also Hermitian) must be smeared out with skew, second-rank test fields. (Note that the right side of (201) is well-defined, for the argument is necessarily divergence-free.)

We next verify that our field operators satisfy the same equations as the classical fields. Using Lemma 1, Eqns. (175) and (176) are translated into

$$\underline{A}(\nabla^a f) = 0 \tag{202}$$

$$\underline{A}(\Box f^a) = 0 \tag{203}$$

where f is any test function and f^a is any test field which is the sum of a gradient and a divergence-free field. Eqn. (202) follows immediately from (199). To prove (203), note that, if $f'^a(k)$ is the Fourier inverse of f^a, then $(-k_b k^b)f'^a(k)$ is the Fourier inverse of $\Box f^a$. But $(-k_b k^b)f'^a(k)$ vanishes on M_0^+, whence (203) follows. We conclude that, in a suitable sense, our vector potential operators satisfy (175) and (176). Similarly, using Lemma 1, Maxwell's equations (170) and (171) on F_{ab} are to be translated into the following conditions on the electromagnetic field operators:

$$\underline{F}(\nabla_c f^{abc}) = 0 \tag{204}$$

$$\underline{F}(\nabla^{[a} f^{b]}) = 0 \tag{205}$$

where f^{abc} is a totally antisymmetric test field and f^a is any test field. To prove (204) and (205), we substitute the definition (201):

$$\underline{F}(\nabla_c f^{abc}) = \underline{A}(2\nabla_b \nabla_c f^{abc}) = \underline{A}(0) \tag{206}$$

$$\underline{F}(\nabla^{[a} f^{b]}) = \underline{A}(2\nabla_b \nabla^{[a} f^{b]}) = \underline{A}(\nabla^a(\nabla_b f^b) - \Box f^a) \tag{207}$$

Thus, (204) is clearly true, while (205) follows immediately from (202) and (203). We conclude that, in a suitable sense, our Maxwell field operators satisfy Maxwell's equations.

Finally, we remark briefly on the commutators of the vector potential operators. Let f^a and $\tilde{f}^a$ be test fields, each of which is the sum of a gradient and a divergence-free field. Then, from (200) and (102),

$$[\underline{A}(f^a), \underline{A}(\tilde{f}^a)] = \hbar^2(-\sigma^\alpha(f^a)\bar{\sigma}_\alpha(\tilde{f}^a) + \sigma^\alpha(\tilde{f}^a)\bar{\sigma}_\alpha(f^a))\mathbb{I} = \frac{\hbar}{i} D(f^a, \tilde{f}^a)\mathbb{I} \tag{208}$$

where the second equality defines $D(f^a, \tilde{f}^a)$. Thus, $D(f^a, \tilde{f}^a)$ is real and satisfies

$$D(f^a, \tilde{f}^a) = -D(\tilde{f}^a, f^a) \tag{209}$$

These properties imply that whenever f^a and $\tilde{f}^a$ have relatively spacelike supports, $D(f^a, \tilde{f}^a) = 0$. As in the Klein-Gordon case, $D(f^a, \tilde{f}^a)$ can be written out explicitly in terms of a distribution on Minkowski space.

15 The Poincaré Group

A smooth mapping from Minkowski space to itself which preserves the norms of vectors is called a *Poincaré transformation*. If, in addition, this mapping i) does not reverse the of past and future time directions, and ii) does not reverse spatial parities (i) and ii) together are equivalent to i) and the condition that ϵ_{abcd} be invariant), then the Poincaré transformation is called a *restricted Poincaré transformation*. The result of applying two Poincaré transformations (resp., restricted Poincaré transformations) in succession is clearly again a Poincaré (resp., restricted Poincaré) transformation. These transformations thus form a group, called the *Poincaré group* (resp. *restricted Poincaré group*), $\mathcal{P}$ (resp., $\mathcal{RP}$.) One sometimes expresses this relation between the Poincaré group and Minkowski space by saying that the Poincaré group *acts* on Minkowski space. That is, we have a mapping $\Psi : \mathcal{P} \times M \to M$ ($M =$ Minkowski space) with the following properties:

$$\Psi(P, \Psi(P', x)) = \Psi(PP', x) \tag{210}$$

$$\Psi(e, x) = x \tag{211}$$

for $P, P' \in \mathcal{P}$, $x \in M$.

In fact, the Poincaré group has more structure than merely that of a group. It is also a (10-dimensional, real, differentiable) manifold. This additional manifold structure on $\mathcal{P}$ leads naturally to the notion of an "infinitesimal Poincaré transformation".

A group G which is also a smooth manifold, and for which the group operations (composition within the group, considered as a mapping from $G \times G$ to G, and the operation of taking the inverse, considered as a mapping from G to G) are smooth mappings, is called a *Lie group*. Let G be a Lie group, and let $\mathcal{L}G$ denote the collection of all contravariant vectors at the identity element e of G. This $\mathcal{L}G$ is thus a real vector space whose dimension is the same as the dimension of the manifold G. (Vectors at the identity of G represent "elements of G which differ infinitesimally from the identity.")

So far, our $\mathcal{L}G$ involves only the manifold structure of G (and, of course, the location of the identity element.) Is there some way in which the group structure of G can also be incorporated into $\mathcal{L}G$? Let $v \in \mathcal{L}G$, so v is a contravariant vector at $e \in G$. Let $g(\epsilon)$ be a smooth curve, parameterized by the parameter ϵ, in G such that $g(O) = e$ and such that the tangent vector, with respect to ϵ, of $g(\epsilon)$ at e is just v. ("Tangent vector with respect to ϵ" means that one takes the derivative of $g(\epsilon)$ with respect to ϵ and evaluates at $\epsilon = 0$.) Similarly, let $g'(\epsilon)$ be a curve associated with $v' \in \mathbb{L}G$. Consider now the smooth curve

$$g(\epsilon)g'(\epsilon)g^{-1}(\epsilon){g'}^{-1}(\epsilon) \tag{212}$$

in G. Unfortunately, the tangent vector (with respect to ϵ) of the curve (212) vanishes at e. It turns out, however, that (212) is still a smooth curve if we take as its parameter not ϵ but rather ϵ^2. The tangent vector of (212), with respect to the parameter ϵ^2, is not in general zero at e. Furthermore, this tangent vector depends only on v and v' (and not on the particular curves $g(\epsilon)$ and $g'(\epsilon)$ which actually appear in (212)), and so we may write it as follows: $[v, v']$. Thus, with any two elements, v and v', of $\mathcal{L}G$, we associate a third element, $[v, v']$, of $\mathcal{L}G$. It is by means of this bracket operation that the group structure of G appears in $\mathcal{L}G$. It can be proven that the bracket is necessarily linear, antisymmetric, and subject to the Jacobi identity:

$$\begin{aligned} [av + bv', v''] &= a[v, v''] + b[v', v''] \\ [v, av' + bv''] &= a[v, v'] + b[v, v''] \end{aligned} \tag{213}$$

$$[v, v'] = -[v', v] \tag{214}$$

$$[v, [v', v'']] + [v', [v'', v]] + [v'', [v, v']] = 0 \tag{215}$$

($a, b \in \mathbb{R}$; $v, v', v'' \in \mathcal{L}G$.) More generally, a *Lie algebra* is a vector space V on which there is given a mapping from $V \times V$ to V (the bracket), subject to (213), (214), and (215).

To summarize, the collection $\mathcal{L}G$ of contravariant vectors at the identity of any Lie group G has the structure of a Lie algebra.

There is a more formal way of expressing the structure of $\mathcal{L}G$ in terms of that of G. Let $v \in \mathcal{L}G$. For each $g \in G$, "left multiplication by G" defines a smooth mapping from G to G which takes e to g. This mapping therefore carries v (a vector at e) to some vector at g. Repeating, for each $g \in G$, we obtain a vector field on G. That is, with each $v \in \mathcal{L}G$ there is associated a certain vector

field on the manifold G. If $v, v' \in \mathcal{L}G$, then the Lie derivative of the vector field associated with v' by the vector field associated with v, evaluated at e, is precisely the element $[v, v']$ of $\mathcal{L}G$. In this formulation, properties (213), (214), and (215) of the bracket are clearly true.

The Poincaré group $\mathcal{P}$ is a Lie group: hence we have a Lie algebra $\mathcal{LP}$. (The Lie algebra of the restricted Poincaré group is the same as the Lie algebra of the Poincaré group: "infinitesimal Poincaré transformations" cannot reverse past and future or spatial parities.) Fortunately, elements of $\mathcal{LP}$ can be expressed very simply as certain vector fields on Minkowski space. This is not surprising: if we think of an element of $\mathcal{LP}$ as representing an "infinitesimal Poincaré transformation," then its action on Minkowski space should be expressible in terms of some vector field on Minkowski space. The vector fields in Minkowski space which represent elements of $\mathcal{LP}$ are those which satisfy

$$\nabla_{(a} v_{b)} = 0 \tag{216}$$

(Eqn. (216) states that the Lie derivative of the Minkowski metric by v^a vanishes.) Choosing a particular origin O, the most general solution of (216) can be expressed in the form

$$v_a = v^O_{ab} x^b + v^O_a \tag{217}$$

where v^O_a is a constant vector field on Minkowski space, v^O_{ab} is a constant skew tensor field on Minkowski space, and x^a is the position vector of x relative to our origin O. Note that the particular constant fields v^O_{ab} and v^O_a which describe a given $v_a(x)$ will depend on the choice of origin O. Note also that the dimensions are correct: six dimensions for v^O_{ab} plus four dimensions for v^O_a make ten dimensions for $\mathcal{LP}$. The bracket operation in $\mathcal{LP}$ becomes Lie derivatives of solutions of (216). That is to say, if $v, v' \in \mathcal{LP}$ correspond to solutions v_a, v'_a, respectively, of (216), then the solution of (216) which corresponds to $[v, v']$ is just

$$\mathcal{L}_v v'^a = v^b \nabla_b v'^a - v'^b \nabla_b v^a \tag{218}$$

As a check, one can verify (213), (214), and (215) for (218).

To summarize, whereas the Lie algebra $\mathcal{LP}$ of the Poincaré group arises from very general considerations involving the structure of Lie groups, $\mathcal{LP}$ can in fact be expressed very simply in terms of certain vector fields in Minkowski space.

16 Representations of the Poincaré Group

Let P be a member of the restricted Poincaré group. Then, with each positive-frequency solution $\phi(x)$ of the Klein-Gordon equation, we may certainly associate another positive-frequency solution, $\phi(Px)$. This mapping from solutions to solutions is clearly linear, and so represent an operator, U_P, on the Hilbert space H_{KG} of positive-frequency solutions of the Klein-Gordon equation. That is, for each $P \in \mathcal{RP}$, we have an operator U_P on H_{KG}. Since the operators arise from the action of $\mathcal{RP}$ on Minkowski space, we have

$$U_P U_{P'} = U_{PP'} \tag{219}$$

$$U_e = \mathbb{I} \tag{220}$$

where e denotes the identity of $\mathcal{RP}$. A mapping from a group into a collection of operators on a Hilbert space, subject to (219) and (220), is called a *representation* of the group. (More generally, the term representation is used when the operators act on any vector space.) Thus, we have defined a representation of $\mathcal{RP}$.

The inner product we have defined on H_{KG} is clearly invariant under the action of the restricted Poincaré group. That is to say, if $P \in \mathcal{RP}$, $\sigma, \tau \in H_{KG}$, we have

$$(U_P\sigma, U_P\tau) = (\sigma, \tau) \tag{221}$$

An operator on a Hilbert space which is defined everywhere and which satisfies (221) for any two elements of that Hilbert space is said to be *unitary*. A representation of a group with the property that the operator associated with each group element is unitary is called a *unitary representation*. We thus have a unitary representation of $\mathcal{RP}$ on H_{KG}.

A similar situation obtains in the Maxwell case (and for the other relativistic field equations we shall introduce later.) We have a unitary representation of $\mathcal{RP}$ on H_M.

Associated with the restricted Poincaré group $\mathcal{RP}$ is its Lie algebra $\mathcal{LP}$. What does a unitary representation of $\mathcal{RP}$ look like in terms of $\mathcal{LP}$? Let U_P be a unitary representation of the restricted Poincaré group on a Hilbert space H. Let $v \in \mathcal{LP}$, and let $P(\epsilon)$ be a corresponding curve in $\mathcal{RP}$. Consider, for each $\sigma \in H$, the right side of

$$H_v\sigma = \frac{\hbar}{i} \lim_{\epsilon \to 0} \frac{U_{P(\epsilon)}\sigma - U_{P(0)}\sigma}{\epsilon} \tag{222}$$

("lim", of course, refers to the topology on H.) It may happen, of course, that the limit in (222) does not exist for certain σ. It is normally the case in practice, however, that the limit does exist for a dense subset of H, and, furthermore, that the limit depends only on v and not on the particular curve $P(\epsilon)$. In this case, the right side of (222) is certainty linear in σ (since the U_P are), and so defines an operator H_v on H. (The factor $\hbar/i$ in (222) is for later convenience.) Thus, we associate with each $v \in \mathcal{LP}$ an operator H_v on H. The operator H_v is linear in v, i.e.,

$$H_{av+bv'} = aH_v + bH_{v'} \tag{223}$$

How is $H_{[v,v']}$ related to H_v and $H_{v'}$? To answer this question, we consider the operators associated with the curve (212):

$$U_{P(\epsilon)}U_{P'(\epsilon)}U^{-1}_{P(\epsilon)}U^{-1}_{P'(\epsilon)} \tag{224}$$

Taking the derivative (i.e., as in (222)) of (224), and evaluating at $\epsilon = 0$, we obtain the desired relation

$$[H_v, H_{v'}] = \frac{\hbar}{i} H_{[v,v']} \tag{225}$$

where we have used (222). In other words, the bracket operation on the v's becomes commutators of the H_v's. (Note that (225) is consistent with (213), (214), and (215).) One further property of the H_v's follows from the unitary character, (221), of our representation. Taking the derivative of

$$(U_{P(\epsilon)}\sigma, U_{P(\epsilon)}\tau) = (\sigma, \tau) \tag{226}$$

with respect to ϵ and evaluating at $\epsilon = 0$, we obtain, using (222),

$$(H_v\sigma, \tau) = (\sigma, H_v\tau) \tag{227}$$

That is, each operator H_v, is Hermitian.

To summarize, a unitary representation of the restricted Poincaré group on a Hilbert space H normally leads to a linear mapping from $\mathcal{LP}$ to the collection of Hermitian operators on H. The Lie bracket operation in $\mathcal{LP}$ translates to the commutator of the corresponding operators.

The general remarks above are merely intended to provide a framework for what follows. In practice, it is not necessary to go through a limiting process to obtain the Hermitian operators associated with a representation of $\mathcal{RP}$. Let $v \in \mathcal{LP}$ be the vector field v_a on Minkowski space, so v_a satisfies (216). Then, if $\phi(x)$ is a positive-frequency solution of the Klein-Gordon equation, so is the right side of

$$H_v\phi = \frac{\hbar}{i} v^a \nabla_a \phi \tag{228}$$

We thus define an operator H_v on (a dense subset of) H_{KG}. The H_v's clearly satisfy (223) and (225). In terms of momentum space, (228) may be described as follows. Let $\phi(k)$ be the Fourier inverse of $\phi(x)$ with respect to an origin O, and let v_a be given by (217) with respect to the same origin O. It then follows immediately, taking the Fourier inverse of (228), that

$$H_v\phi(k) = \hbar(v^{0a}k_a)\phi(k) + \frac{\hbar}{i} v^{0a}{}_b k^b \partial_a \phi(k) \tag{229}$$

(Note that (229) is well-defined, for $v^{0a}{}_b k^b$ tangent to M_μ^+.) We see from (229) and (30) that each H_v is Hermitian.

The situation is completely analogous for the Maxwell Hilbert space H_M. Eqn. (228) is replaced by

$$H_v F_{ab} = \frac{\hbar}{i}\mathcal{L}_v F_{ab} = \frac{\hbar}{i}(v^c \nabla_c F_{ab} + F_{cb}\nabla_a v^c + F_{ac}\nabla_b v^c) \tag{230}$$

where the multiplication by i in (230) refers to multiplication within the Hilbert space H_M. In momentum space, our Hermitian operators take the form

$$H_v A'^a(k) = \hbar(v^{0b}k_b)A'^a(k) + \frac{\hbar}{i}\mathcal{L}_{v^{0c}{}_b k^b} A'^a(k) \tag{231}$$

To summarize, we can take $\mathcal{LP}$ be simply the Lie algebra of solutions of (216), and the operators H_v, to be defined by (228) and (230) (or by (229) and (231)). Then Hermiticity, (223), and (225) follow directly.

To facilitate calculations with the H_v's, it is convenient to introduce a special notation. Let $T^{a_1\cdots a_n}$ be a tensor field on Minkowski space. Then $T^{a_1\cdots a_n}$ associates, with each point x and tensor $f_{a_1\cdots a_n}$ at x, a real number, $T^{a_1\cdots a_n}f_{a_1\cdots a_n}$. For fixed x, this mapping is linear in $f_{a_1\cdots a_n}$. Furthermore, the value of this number for every x and $f_{a_1\cdots a_n}$ determines $T^{a_1\cdots a_n}$ uniquely. (Think of $f_{a_1\cdots a_n}$ as a "test function.") An operator field is what results if we replace "real number" in the remarks above by "operator on a Hilbert space H." Thus, an *operator field*, $\underline{T}^{a_1\cdots a_n}$, associates, with each point x of Minkowski space and tensor $f_{a_1\cdots a_n}$ at x, an operator on H, written $\underline{T}^{a_1\cdots a_n}f_{a_1\cdots a_n}$, such that, for x fixed, this operator is linear in $f_{a_1\cdots a_n}$. (For example, an operator field is what $\underline{A}_a(x)$ and $\underline{F}_{ab}(x)$ would be, if they existed.) Note that a tensor field is a special case of an operator field — when all the operators are multiples of the identity operator on H.

The easiest way to discuss the H_v's is as operator fields. Let x be a point of Minkowski space, and f_a a vector at x. Then the constant vector field

$$v_a = f_a \tag{232}$$

on Minkowski space certainly satisfies (216), and so defines an operator H_v (on either H_{KG} or H_M). We have defined an operator field, which we write as $\underline{P}_a$. (These, of course, are our old momentum operators, expressed in a different way.) Let x be point of Minkowski space, and let f_{ab} be a skew tensor at x. Then the vector field

$$v_a(y) = f_{ab}x^b \tag{233}$$

on Minkowski space, where x^a denotes the position vector of y relative to x, satisfies (216), and so defines an operator H_v. We have thus defined a skew operator field, which we write P_{ab}.

We introduce three operations on operator fields. The first is outer product. Let f^{abc} be a tensor at the point x of Minkowski space. Write f^{abc} in the form

$$f^{abc} = m^a m^{bc} + \cdots + n^a n^{bc} \tag{234}$$

Then, for example, the outer product of $\underline{P}_a$ and $\underline{P}_{bc}$ is the operator field $\underline{P}_a\underline{P}_{bc}$, defined by

$$\underline{P}_a\underline{P}_{bc}f^{abc}a = (m^a\underline{P}_a)(m^{bc}\underline{P}_{bc}) + \cdots (n^a\underline{P}_a)(n^{bc}\underline{P}_{bc}) \tag{235}$$

where the products on the right are to be interpreted as merely products of operators. Note that (235) is independent of the particular expansion (234). The outer product of two operator fields in general depends on the order in which they are written. For example, $\underline{P}_a\underline{P}_{bc} \neq \underline{P}_{bc}\underline{P}_a$. The second operation is contraction. Let f^b be a vector at the point x of Minkowski space. Then, for example, $\underline{P}^a\underline{P}_{ab}$ is the operator defined by

$$\begin{aligned}\underline{P}^a\underline{P}_{ab}f^b = (\underline{P}^c t_c)(\underline{P}_{db}t^d f^b) - (\underline{P}^c x_c)(\underline{P}_{db}x^d f^b) - (\underline{P}^c y_c)(\underline{P}_{db}y^d f^b) \\ - (\underline{P}^c z_c)(\underline{P}_{db}z^d f^b)\end{aligned} \tag{236}$$

where t^a, x^a, y^a, z^a are vectors at x which define an orthonormal basis

$$\begin{gathered} t^a t_a = 1 \qquad x^a x_a = y^a y_a = z^a z_a = -1 \\ t^a x_a = t^a y_a = t^a z_a = x^a y_a = x^a z_a = y^a z_a = 0 \end{gathered} \tag{237}$$

Note that (236) is independent of the choice of basis. The final operation on operator fields is differentiation. Let r^a and f^b be vectors at the point x of Minkowski space. Let x' be the point of Minkowski space whose position vector relative to x is ϵr^a, and let f'^b be f^b translated to the point x'. Then, for example, $\nabla_a \underline{P}_b$ is the operator field defined by

$$(\nabla_a \underline{P}_b) r^a f^b = \lim_{\epsilon \to 0} \frac{\underline{P}_a f'^a - \underline{P}_a f^a}{\epsilon} \tag{238}$$

(provided this limit exists). In short, operator fields are handled exactly as tensor fields, except that one must keep track of the order in products. The terms Hermitian operator field, unitary operator field, etc. are self-explanatory.

First note that $\underline{P}_a$ and $\underline{P}_{ab}$ are Hermitian operator fields. We next consider the derivatives of our two operator fields. It is clear from (232) and (238) that $\underline{P}_a$ is constant:

$$\nabla_a \underline{P}_b = 0 \tag{239}$$

To compute the derivative $\underline{P}_a b$, we first note the following fact. If $v_a(x)$, a solution of (216), is expressed in the form (217) with respect to two different origins, O and O', then

$$v_{ab}^{O'} = v_{ab}^{O} \qquad v_a^{O'} = v_a^O + v_{ab}^O r^b \tag{240}$$

where r^a is the position vector of O' relative to O. It now follows from (233) and (240) that, if s^{cd} is a skew tensor at O and s'^{cd} is s^{cd} translated to O', then

$$\underline{P}_{cd} s'^{cd} - \underline{P}_{cd} s^{cd} = -s_{cd} r^d \underline{P}^c \tag{241}$$

Hence,

$$\nabla_a \underline{P}_{bc} = \eta_{a[b} \underline{P}_{c]} \tag{242}$$

Eqns. (239) and (242) imply, in particular, that the second derivative of $\underline{P}_a b$ vanishes. Finally, we evaluate the commutators of our operator fields. We have already seen that the momentum operators commute:

$$[\underline{P}_a, \underline{P}_b] = 0 \tag{243}$$

The other commutators are computed using the following fact: if $v_a(x)$ and $w_a(x)$ are elements of $\mathcal{LP}$, expressed in the form (217) with respect to the same origin O, then $[v, w]$ takes the form

$$2 v_{[a}^{Oc} w_{b]c}^O x^b + (v^O c w_{ac}^O - w^{Oc} v_{ac}^O) \tag{244}$$

with respect to O. Hence, from (225), (233), and (244), we have

$$[r^{ab} \underline{P}_{ab}, s^{cd} \underline{P}_c d] = \frac{2\hbar}{i} r_a{}^c s_{bc} \underline{P}^{ab} \tag{245}$$

where r^{ab} and s^{ab} are skew tensors at x. Therefore,

$$[\underline{P}_{ab}, \underline{P}_{cd}] = \frac{\hbar}{i}\left(\eta_{b[c}\underline{P}_{d]a} - (\eta_{a[c}\underline{P}_{d]b}\right) \tag{246}$$

By an identical argument, we obtain, finally,

$$[\underline{P}_{a}, \underline{P}_{bc}] = -\frac{\hbar}{i}\eta_{a[b}\underline{P}_{c]} \tag{247}$$

The interaction between the restricted Poincaré group and our Hilbert spaces is expressed completely and neatly by the operator fields $\underline{P}_a$ and $\underline{P}_{ab}$. The important equations on these fields are (239), (242), (243), (246), and (247).

17 Casimir Operators: Spin and Mass

Our plan is to introduce a number of relativistic field equations, and, for each one, to make a Hilbert space of an appropriate collection of solutions, to introduce the corresponding Fock space, and to replace the classical fields by operators on Fock space. This program has now been carried out for the Klein-Gordon and Maxwell equations. With each set of equations there are associated two real numbers called the mass and the spin. We could, of course, merely state what mass and what spin are to be associated with the equations in each case. It is useful, however, to see how these quantities arise in a natural way from very general considerations involving the structure of the Poincaré group. In fact, what we need of the Poincaré group is the action of its Lie algebra, $\mathcal{LP}$, on our Hilbert spaces (Sect. 16), and certain objects, called Casimir operators, associated with $\mathcal{LP}$. More generally, there are Casimir operators associated with any Lie algebra. We begin with this more general situation.

Let $\mathcal{L}$ be a Lie algebra. Then, in particular, $\mathcal{L}$ is a vector space. It is convenient to introduce an index notation. An element of $\mathcal{L}$ will be written with a raised Greek index (not to be confused with the Greek indices used in the discussion of Fock space.) Elements of the dual space of $\mathcal{L}$ (elements of the vector space of linear maps from $\mathcal{L}$ to the reals (or the complexes, if $\mathcal{L}$ were a complex vector space)) are written with lowered Greek indices. Objects with more that one index represent tensors over $\mathcal{L}$ and its dual. Finally, the action of the dual induces the operation of contraction between one raised and one lowered Greek index: this is indicated by using a repeated index. (When one wants to do anything except the most trivial calculations with multilinear algebra, it is usually simpler in the long run to introduce an index notation.) For example, the bracket operation in $\mathcal{L}$ is a bilinear mapping from $\mathcal{L} \times \mathcal{L}$ to $\mathcal{L}$, and so can be represented by a tensor $C^{\mu}{}_{\alpha\beta}$:

$$[v, v']^{\mu} = C^{\mu}{}_{\alpha\beta} v^{\alpha} v'^{\beta} \tag{248}$$

(This tensor is sometimes called the *structure constant tensor.*) Eqns. (214) and (215), expressed in terms of $C^{\mu}{}_{\alpha\beta}$, become

$$C^{\mu}{}_{\alpha\beta} = C^{\mu}{}_{[\alpha\beta]} \tag{249}$$

$$C^{\mu}{}_{\nu[\alpha}C^{\nu}{}_{\beta\gamma]} = 0 \tag{250}$$

In other words, a Lie algebra is simply a vector space over which there is given a tensor $C^{\mu}{}_{\alpha\beta}$, subject to (249) and (250). (For example, the invariant metric of a Lie algebra, which we shall not use here, is $C^{\mu}{}_{\nu\alpha}C^{\nu}{}_{\mu\beta}$.)

We now introduce the set $\mathcal{A}$ of all finite strings of tensors over $\mathcal{L}$:

$$(v^{\alpha}, v^{\alpha}\beta, \ldots, v^{\alpha_1\cdots\alpha_n}, 0, 0, \ldots) \tag{251}$$

What structure do we have on $\mathcal{A}$? We can certainly add two finite strings by adding them "component-wise" — i.e., adding the vector of the first string to the vector of the second string, the second-rank tensor of the first string to the second-rank tensor of the second string, etc. — to obtain a new element of $\mathcal{A}$. Furthermore, we can multiply a finite string by a number by multiplying each element of that string by the number. Thus, $\mathcal{A}$ has the structure of an (infinite-dimensional) vector space. We can also introduce a product operation on $\mathcal{A}$. (This, in fact, is the reason for considering $\mathcal{A}$ at all.) To take the product of two finite strings, take all possible outer products consisting of one tensor from the first string and one from the second, always placing the tensor from the first string first, and add together the resulting tensors when they have the same rank to obtain the product string. For example,

$$\begin{aligned}(v^{\alpha}, v^{\alpha\beta}, 0, \ldots)(w^{\alpha}, w^{\alpha\beta}, w^{\alpha\beta\gamma}, 0, \ldots)\\ = (0, v^{\alpha}w^{\beta}, v^{\alpha\beta}w^{\gamma} + v^{\alpha}w^{\beta\gamma}, v^{\alpha\beta}w^{\gamma\delta} + v^{\alpha}w^{\beta\gamma\delta}, v^{\alpha\beta}w^{\gamma\delta\epsilon}, 0, \ldots)\end{aligned} \tag{252}$$

Note that the product, AB, of elements A and B of $\mathcal{A}$ is linear in A and B:

$$\begin{aligned}(aA + A')B &= aAB + A'B\\ A(aB + B') &= aAB + AB'\end{aligned} \tag{253}$$

($a \in \mathbb{R}$, $A, A', B, B' \in \mathcal{A}$.) A vector space on which there is defined a product which satisfies (253) is called an *algebra*. So $\mathcal{A}$ is an algebra. (Note that every Lie algebra is an algebra: the product is the bracket.) Since outer products of tensors are associative, so is $\mathcal{A}$:

$$A(BC) = (AB)C \tag{254}$$

($A, B, C \in \mathcal{A}$). An algebra for which (254) holds for any three of its elements is called an *associative algebra*.

Our algebra $\mathcal{A}$ so far involves only the vector-space structure of $\mathcal{L}$. (In fact, $\mathcal{A}$ is sometimes called the tensor algebra of the vector space $\mathcal{L}$.) We now want to incorporate in some way the remaining structure of $\mathcal{L}$, i.e., the bracket. Consider the collection of all elements of $\mathcal{A}$ of the form

$$(-C^{\alpha}{}_{\mu\nu}v^{\mu}w^{\nu}, 2v^{[\alpha}w^{\beta]}, 0, \ldots) \tag{255}$$

for $v, w \in \mathcal{L}$. Let $\mathcal{I}$ denote the set of all elements of $\mathcal{A}$ which can be written as a sum of products of elements of $\mathcal{A}$ in such a way that at least one factor in

each product is of the form (255). Clearly, we have (i) $\mathcal{I}$ is a vector subspace of $\mathcal{A}$, and (ii) the product of any element of $\mathcal{I}$ with any element of $\mathcal{A}$ is again an element of $\mathcal{I}$. (A subset of an associative algebra, satisfying (i) and (ii), is called an ideal.) We now want to take the quotient algebra, $\mathcal{A}/\mathcal{I}$, of $\mathcal{A}$ by the ideal $\mathcal{I}$. We define an equivalence relation on $\mathcal{A}$: two elements of $\mathcal{A}$ are to be regarded as equivalent if their difference is in $\mathcal{I}$. That the equivalence class of any linear combination of elements A and B of $\mathcal{A}$ depends only on the equivalence classes of A and B follows from (i). That the equivalence class of the product of any two elements A and B of $\mathcal{A}$ depends only on the equivalence class of A and B follows from (ii). Thus, the collection of equivalence classes is itself an associative algebra. It is written $\mathcal{UL}$ and called the *universal enveloping algebra* of the Lie algebra. To summarize, with every algebra there is associated an associative algebra $\mathcal{UL}$.

There is an important relation between $\mathcal{L}$ and $\mathcal{UL}$. Let $v^\alpha \in \mathcal{L}$, and let $\psi(v)$ denote the element of $\mathcal{UL}$ whose equivalence class contains the element $(v^\alpha, 0, 0, \ldots)$ of $\mathcal{A}$. We thus have a — clearly linear — mapping from to $\mathcal{L}$ to $\mathcal{UL}$. Furthermore, it follows from (255) that

$$\psi([v, v']) = \psi(v)\psi(v') - \psi(v')\psi(v) \tag{256}$$

for any two elements of $\mathcal{L}$. (In fact, it was to make (256) hold that we defined $\mathcal{UL}$ as we did.) In other words, the bracket operation in the Lie algebra $\mathcal{L}$ corresponds to the commutator of elements of the associative algebra $\mathcal{UL}$. Note that, applying ψ to both sides of (214) and (215), and using (256) and associativity, we obtain identities.

Why this interest in the universal enveloping algebra? Let $\mathcal{L}$ be a Lie algebra, and suppose, for each element v of $\mathcal{L}$, we are given an operator H_v on some fixed Hilbert space H. Suppose, furthermore, that H_v is linear in v, and that

$$H_{[v,v']} = H_v H_{v'} - H_{v'} H_v \tag{257}$$

for any $v, v' \in \mathcal{L}$. (Compare (225). It is convenient to omit the factors $\hbar/i$ — not an essential change, for such factors can always be included in the operators — when considering purely mathematical questions.) We show that this action of $\mathcal{L}$ on H can be extended naturally to an action of $\mathcal{UL}$ on H. Consider an element of the associative algebra $\mathcal{A}$, written (as every element of $\mathcal{A}$ can be written) as sums of outer products of vectors, e.g.

$$(v^\alpha, u^\alpha w^\beta + p^\alpha q^\beta, r^\alpha s^\beta t^\beta, 0, 0, \ldots) \tag{258}$$

We associate with each expression of the form (258) an operator on H, e.g.,

$$H_v + H_u H_w + H_p H_q + H_r H_s H_t \tag{259}$$

It follows from the fact that H_v is linear in v that the operator (259) depends only on the element of $\mathcal{A}$ represented by (258) (and not on the particular expansion used.) Furthermore, (255) and (257) imply that if (258) is an element of $\mathcal{I}$ ($\in \mathcal{A}$), then the operator (259) is zero. Thus, (259) depends only on the

equivalence class of (258). In other words, we have, for each element Γ of $\mathcal{UL}$, an operator, H_Γ, on H. The operators H_Γ are linear in H:

$$H_{a\Gamma+\Gamma'} = aH_\Gamma + H_{\Gamma'} \tag{260}$$

and, clearly, satisfy

$$H_{\psi(v)} = H_v \tag{261}$$

Furthermore, it follows immediately from (259) that

$$H_{\Gamma\Gamma'} = H_\Gamma H_{\Gamma'} \tag{262}$$

Let us summarize the situation. We have a Lie algebra $\mathcal{L}$ acting on a Hilbert space H by means of the operators H_v ($v \in \mathcal{L}$) on H. The collection of all operators (at least, the collection of all those which are defined everywhere) on a Hilbert space has the structure of an associative algebra. We thus have a mapping from a Lie algebra to an associative algebra, with these two algebraic structures related via (257). Things could be better. It would be nice if we could express the bracket operation in $\mathcal{L}$ in the form

$$[v, v'] = vv' - v'v \tag{263}$$

and have

$$H_{vv'} = H_v H'_v \tag{264}$$

Then (257) would follow already from (263) and (264). This program, unfortunately, cannot be accomplished directly, for the only "product" which is defined in $\mathcal{L}$ is the entire bracket, and not the individual terms on the right of (263). But it can be accomplished indirectly. We "enlarge" $\mathcal{L}$ to $\mathcal{UL}$. We still cannot write (263) — but instead we have (256). (Eqn. (256) also states that the algebraic structure of $\mathcal{L}$ has been incorporated into that of $\mathcal{UL}$.) We still cannot write (264) — but instead we have (262). In short, since $\mathcal{L}$ is being mapped to an associative algebra (the operators on H), and since the natural thing to map to an associative algebra is another associative algebra, we "force associativity" on $\mathcal{L}$ by enlarging it to $\mathcal{UL}$.

We can now introduce the Casimir operators. A *Casimir operator* of the Lie algebra $\mathcal{L}$ is an element of the center of $\mathcal{UL}$, i.e., an element Γ of $\mathcal{UL}$ such that

$$\Gamma\Delta - \Delta\Gamma = 0 \tag{265}$$

for every $\Delta \in \mathcal{UL}$. It should be emphasized that the Casimir operators of a Lie algebra are not themselves elements of that Lie algebra, but rather of its universal enveloping algebra. That is, they must be represented as strings of tensors over $\mathcal{L}$. Note that the collection of all Casimir operators of a Lie algebra form an associative algebra. Finally, we remark that the universal enveloping algebra $\mathcal{UL}$ and hence the Casimir operators (which are not operators, as we have defined them, but merely elements of an algebra) are fixed once and for all given the Lie algebra $\mathcal{L}$. They do not depend on the presence of a Hilbert space

H or on the H_v's. For example, the Casimir operators of $\mathcal{LP}$ (the Lie algebra of the Poincaré group) simply exist. (In fact, there are just two algebraically independent ones.) They needn't be found individually for H_{KG}, H_M, etc.

Now suppose again that we have a Hilbert space H and, for each $v \in \mathcal{L}$, an operator H_v on H, where the H_v's are linear in v and satisfy (257). Then we have an operator H_Γ on H for each $\Gamma \in \mathcal{UL}$, and, in particular, an operator on H for each Casimir operator of $\mathcal{L}$. Eqns. (265) and (262) imply that, if Γ is a Casimir operator, then H_Γ commutes with all the H_Δ's, and, in particular, with all the H_v's (see (261)). This normally implies, as we shall see in examples later, that H_Γ is simply a multiple of the identity operator. Thus, the Casimir operators assign numbers to our relativistic field equations, i.e., to our representations of the restricted Poincaré group.

The words "normally implies" above are rather vague. I do not know whether or not there is a general theorem which implies that the Casimir H_Γ's are multiples of the identity in cases of interest. However, the following result suggests this conclusion:

Lemma 2 (Schur's Lemma). *Let H be a finite-dimensional complex vector space, and let $\mathcal{L}$ be a set. Suppose, for each $v \in \mathcal{L}$, we are given an operator (defined everywhere) on H, H_v. Suppose, furthermore, that the only vector subspaces S of H having the property that $H_v\sigma \in S$ for every $v \in \mathcal{L}$ and $\sigma \in S$ are $S = \{0\}$ and $S = H$. Let K be an operator (defined everywhere) on H which commutes with all the H_v's. Then K is some complex multiple of the identity.*

Proof. Since H is a complex vector space, K has at least one eigenvector, i.e., there exists a complex number κ and a nonzero element σ of H such that

$$K\sigma = \kappa\sigma \tag{266}$$

Fix κ, and let S be the collection of all σ's which satisfy (266). Then, for $\sigma \in S$, $v \in \mathcal{L}$,

$$K(H_v\sigma) = H_vK\sigma = \kappa(H_v\sigma) \tag{267}$$

Hence, $H_v\sigma \in S$. By hypothesis, therefore, $S = \{0\}$, or $S = H$. But by construction S contains at least one nonzero element of H, so we must have $S = H$. In other words, every element of H satisfies (266), whence $K = \kappa I$. □

We now want to apply all this mathematics to our relativistic fields. As usual, one can regard the formal developments as merely providing motivation and insight into what turn out to be very simple notions in practice. The operators on our Hilbert spaces associated with the Casimir operators of $\mathcal{LP}$ can be expressed quite easily in terms of the operator fields $\underline{P}_a$ and $\underline{P}_{ab}$ discussed in Section 16. The first Casimir operator is

$$\underline{P}_a\underline{P}^a = m^2 \tag{268}$$

We see from (243) and (247) that m^2 commutes with $\underline{P}_a$ and $\underline{P}_{ab}$. Furthermore, (239)) implies that m^2 is a constant operator field. Hence, m^2 is just an ordinary

operator on our Hilbert spaces. It turns out to be a multiple of the identity (as suggested above), and that multiple is called the (squared) *mass* of the field. To define the second Casimir operator, we first introduce the operator field

$$\underline{W}^a \epsilon^{abcd} \underline{P}_b \underline{P}_{cd} \tag{269}$$

Then (239) and (242) imply that $\underline{W}^a$ is constant. The second Casimir operator is the left side of

$$\underline{W}_a \underline{W}^a = -\hbar^2 m^2 s(s+1) \tag{270}$$

Note, from (243), (246), and (247), that $\underline{W}_a \underline{W}^a$ commutes with $\underline{P}_a$ and $\underline{P}_{ab}$. It turns out to be a multiple of the identity, and the non-negative number s which makes that multiple be the right side of (270) is called the *spin* of the field. We remark that the mass and spin are associated not with each individual solution of a relativistic field equation, but rather with the equation itself.

Unfortunately, (270) will not give the spin s when $m = 0$. In the massless case, it is found that there is a number s for which

$$\underline{W}_a = s\hbar \underline{P}_a \tag{271}$$

and so this equation is used to define the spin. This definition has an interesting consequence. Note that the definition of $\underline{W}^a$ involves one ϵ^{abcd}, while there are none in $\underline{P}_a$. That is, $\underline{W}^a$ is a pseudovector, while $\underline{P}^a$ an ordinary vector (operator field). Hence, the spin s is a pseudoscalar in the massless case, and a scalar when $m \neq 0$. We shall see shortly that this feature is related to the notion of helicity.

Finally, we evaluate the mass and spin in the Klein-Gordon and Maxwell cases. Let r^a be a vector at the point x of Minkowski space, and let positive-frequency solution of the Klein-Gordon equation. Then

$$r^a \underline{P}_a (r^b \underline{P}_b \phi) = r^a \underline{P}_a \left(\frac{\hbar}{i} r^b \nabla_b \phi \right) = -\hbar^2 r^a r^b \nabla_a \nabla_b \phi \tag{272}$$

To evaluate $\underline{P}_a \underline{P}^a \phi$ we must sum (272), with the appropriate signs, as r^a runs over an orthonormal tetrad (see (236).) Clearly, the result of taking this sum is simply to replace $r^a r^b$ by the Minkowski metric, η^{ab}. So

$$\underline{P}_a \underline{P}^a \phi = -\hbar^2 \Box \phi = \hbar^2 \mu^2 \phi \tag{273}$$

But $\hbar^2 \mu^2$ for the Klein-Gordon equation is what we earlier (c.f. (5)) called m^2. Hence, the m in Sect. 1 is indeed the mass for the Klein-Gordon equation. To evaluate the spin, let r^a be a vector and s^{ab} a skew tensor at the point x. Then, writing x^a for the position vector relative to x,

$$\begin{aligned} r^a \underline{P}_a (s^{bc} \underline{P}_{bc} \phi) = r^a \underline{P}_a \left(\frac{\hbar}{i} s^b{}_c x^c \nabla_b \phi \right) = \frac{\hbar}{i} r^a \nabla_a \left(\frac{\hbar}{i} s^b{}_c x^c \nabla_b \phi \right) \\ = -\hbar^2 (r^c s^b{}_c \nabla_b \phi + r^a s^b{}_c x^c \nabla_a \nabla_b \phi) \end{aligned} \tag{274}$$

Let u^a be another vector at x. Then, to evaluate $u_a \underline{W}^a \phi$, we must sum (274) over r's and s's so that $\sum r^b s^{cd} = u_a \epsilon^{abcd}$. The result, clearly, is just to replace the combination $r^b s^{cd}$ in (274) by $u_a \epsilon^{abcd}$. So,

$$u_a \underline{W}^a \phi = -\hbar^2 u_a \epsilon^{abcd} (\eta_{bd} \nabla_c \phi + x_d \nabla_b \nabla_c \phi) = 0 \tag{275}$$

Thus, $\underline{W}^a$ is zero on H_{KG}. Now (270) implies $s = 0$ in the massive case, while (271) gives $s = 0$ in the massless case. The Klein-Gordon equation describes a particle of mass m and spin zero.

It is enlightening, instead of treating just the Maxwell case, to discuss the more general equation

$$(\Box + \mu^2) A_a = 0 \qquad\qquad \nabla^a A_a = 0 \tag{276}$$

Maxwell's equations are obtained for $\mu = 0$. If r^a is a vector at x,

$$r^a \underline{P}_a (r^b \underline{P}_b A_c) = -\hbar^2 r^a r^b \nabla_a \nabla_b A_c \tag{277}$$

Hence,

$$\underline{P}_a \underline{P}^a A_b = -\hbar^2 \Box A_b = \hbar^2 \mu^2 A_b \tag{278}$$

Hence, the mass of the fields described by (276) is just as in the Klein-Gordon case. (In particular, photons have mass zero.) If r^a is a vector and s^{ab} a skew tensor at x, then

$$\begin{aligned} r^a \underline{P}_a (s^{bc} \underline{P}_{bc} A_d) &= r^a \underline{P}_a \left(\frac{\hbar}{i} \mathcal{L}_{s^b{}_c x^c} A_d \right) \\ &= \frac{\hbar}{i} r^a \nabla_a \left(\frac{\hbar}{i} \left(s^b{}_c x^c \nabla_b A_d + A_b s^b{}_d \right) \right) \\ &= -\hbar^2 \left[r^a s^b{}_c x^c \nabla_a \nabla_b A_d + r^c s^b{}_c \nabla_b A_d + r^a s^b{}_d \nabla_a A_b \right] \end{aligned} \tag{279}$$

Therefore, by the same argument as before,

$$\begin{aligned} u_a \underline{W}^a A_e &= -\hbar^2 u_a \epsilon^{abcd} \left[x_d \nabla_b \nabla_c A_e + \eta_{bd} \nabla_c A_e + \eta_{ed} \nabla_b A_c \right] \\ &= -\hbar^2 \epsilon_{aebc} u^a \nabla^b A^c \end{aligned} \tag{280}$$

where u_a is a vector at x. Hence,

$$\begin{aligned} \underline{W}_a \underline{W}^a A_c &= -\hbar^2 \epsilon_{aebc} \nabla^b \left(-\hbar^2 \epsilon^{acpq} \nabla_p A_q \right) \\ &= -4\hbar^4 \nabla^b (\nabla_{[e} A_{b]}) = -2\hbar^4 \mu^2 A_e \end{aligned} \tag{281}$$

Thus, the spin of the fields (276) is $s = 1$, provided $\mu \neq 0$.

But something appears to be wrong in the Maxwell case, $\mu = 0$. Eqn. (280) is not proportional to

$$u_a \underline{P}^a A_e = \frac{\hbar}{i} r^a \nabla_a A_e \tag{282}$$

First note that, by a gauge transformation on the right in (282), we can write

$$u_a \underline{P}^a A_e = \frac{\hbar}{2i} u^a \nabla_{[a} A_{e]} \tag{283}$$

We still don't have proportionality with (280). The reason is that the representation of $\mathcal{LP}$ on H_M is not irreducible. A solution of Maxwell's equations is said to have *positive* (resp. *negative*) *helicity* if

$$\epsilon^{abcd} F_{cd} = \pm \frac{i}{2} F^{ab} \tag{284}$$

with the plus (resp. minus) sign on the right. (In (284), "i" means multiplication in H_M. The factor $i/2$ is necessary because $\epsilon^{abcd}\epsilon_{cdef}F^{ef} = -4F^{ab}$ for any skew F^{ab}.) In momentum space, a positive-helicity or negative-helicity solution takes the form

$$A'_a(k) = m_a + i n_a \tag{285}$$

with $m^a m_a = n^a n_a$, $m^a n_a = 0$. The two helicities arise because there are two directions through which m_a can be rotated through 90^o to obtain n_a. Every solution of Maxwell's equations can be written uniquely as the sum of a positive and a negative helicity solution. Furthermore, the inner product of a positive helicity solution with a negative helicity solution is zero. (These facts follow immediately from (285).) Thus, H_M is the direct sum of the Hilbert space of positive-helicity solutions with the Hilbert space of negative-helicity solutions. On the Hilbert space of positive-helicity solutions, $s = 1$; on the Hilbert space of negative-helicity solutions, $s = -1$.

18 Spinors

Particles with half-integer spin (electrons, neutrinos, etc.) are described by mathematical objects called spinor fields. We shall base our treatment of such particles on what are called "two-component spinors" (rather than the more common four-component spinors.) Essentially the only difference between the two is one of notation. Whereas the two-component spinors lend themselves more naturally to an index notation, the four-component spinors are slightly more convenient when discussing discrete symmetries. We shall first define (two-component) spinors, and then indicate how formulae can be translated to the four-component language.

Let $\mathfrak{C}$ be a two-dimensional, complex vector space. Membership in $\mathfrak{C}$ will be indicated with a raised, upper case Latin index, e.g., ξ^A, η^A, etc. We introduce three additional two-dimensional complex vector spaces:

i) The complex-conjugate space, $\bar{\mathfrak{C}}$, of $\mathfrak{C}$ (see Sect. 8);

ii) The dual space, $\mathfrak{C}^*$, of $\mathfrak{C}$ (i.e., the vector space of linear mappings from $\mathfrak{C}$ to the complexes; and

iii) The complex-conjugate space of the dual space of $\mathfrak{C}$, $\bar{\mathfrak{C}}^*$ (or, what is the same thing, the dual space of the complex-conjugate space of $\mathfrak{C}$).

Membership in $\bar{\mathfrak{C}}$ will be indicated with a primed raised index, e.g., $\xi^{A'}$; membership in $\mathfrak{C}^*$ by an unprimed lowered index, e.g., ξ_A; and membership in $\bar{\mathfrak{C}}^*$ by a primed lowered index, e.g., $\xi_{A'}$. That is, we now have four different vector spaces, with their elements represented by four different index combinations. What operations can be performed on the elements of $\mathfrak{C}$, $\bar{\mathfrak{C}}$, $\mathfrak{C}^*$, and $\bar{\mathfrak{C}}^*$? Of course, we can multiply elements by complex numbers, and add elements which belong to the same vector space (i.e., which have the same index structure). Furthermore, our four vector spaces can be grouped into pairs which are complex-conjugates of each other: $\mathfrak{C}$ and $\bar{\mathfrak{C}}$ are complex-conjugates of each other, and $\mathfrak{C}^*$ and $\bar{\mathfrak{C}}^*$ are complex-conjugates of each other. Thus, we have an operation of complex-conjugation, which, applied to an element of $\mathfrak{C}$ (resp. $\bar{\mathfrak{C}}$, $\mathfrak{C}^*$, or $\bar{\mathfrak{C}}^*$), yields an element of $\bar{\mathfrak{C}}$ (resp. $\mathfrak{C}$, $\bar{\mathfrak{C}}^*$, $\mathfrak{C}^*$). For example,

$$\begin{aligned} \overline{\xi^A} &= \bar{\xi}^{A'} & \qquad \overline{\eta^{A'}} &= \bar{\eta}^A \\ \overline{\sigma_A} &= \bar{\sigma}_{A'} & \qquad \overline{\tau_{A'}} &= \bar{\tau}_A \end{aligned} \tag{286}$$

Note the effect of the operation of complex-conjugation on the index structure: it adds a prime if there was none before, and a deletes a prime if there was one before. Finally, we can group our four vector spaces into pairs which are duals of each other: $\mathfrak{C}$ and $\mathfrak{C}^*$ are duals of each other, and $\bar{\mathfrak{C}}$ and $\bar{\mathfrak{C}}^*$ are duals of each other. We thus have the operation of contraction: an element $\xi^A \in \mathfrak{C}$ together with an element $\eta_A \in \mathfrak{C}^*$ defines a complex number, $\xi^A\eta_A$; an element $\sigma^{A'} \in \bar{\mathfrak{C}}$ together with an element $\tau_{A'} \in \bar{\mathfrak{C}}^*$ defines a complex number, $\sigma^{A'}\tau_{A'}$. One indicates contraction, as above, by using a repeated index. Note that one can only contract between a raised and a lowered index when these are of the same type (both primed or both unprimed). We have, for example,

$$\overline{(\xi^A\eta_A)} = \bar{\xi}^{A'}\bar{\eta}_{A'} \tag{287}$$

(Note that the index notation we used for Hilbert spaces is essentially a special case of that described above. The inner product on a Hilbert space induces a natural isomorphism between $\bar{\mathfrak{C}}$ and $\mathfrak{C}^*$, and between $\bar{\mathfrak{C}}^*$ and $\mathfrak{C}$. One can therefore do away with primed indices entirely.)

Now consider the various tensor products between $\mathfrak{C}$, $\bar{\mathfrak{C}}$, $\mathfrak{C}^*$, and $\bar{\mathfrak{C}}^*$. The particular tensor product to which an object belongs is indicated by its index structure, e.g., $T^{A\cdots BC'\cdots D'}{}_{E\cdots FG'\cdots H'}$. Complex-conjugation extends in an obvious way to the tensor products, e.g.,

$$\overline{T^{A\cdots BC'\cdots D'}{}_{E\cdots FG'\cdots H'}} = \bar{T}^{A'\cdots B'C\cdots D}{}_{E'\cdots F'G\cdots H} \tag{288}$$

We define a *spinor-space* as a two-dimensional, complex vector space $\mathfrak{C}$ on which is specified a nonzero object ϵ_{AB} which is skew:

$$\epsilon_{AB} = \epsilon_{[AB]} \tag{289}$$

(Note that, since $\mathfrak{C}$ is two-dimensional, any two skew ϵ_{AB}'s differ at most by a complex factor. Hence, there is "just one" spinor space.) Elements of the tensor products will be called *spinors*. Thus, we can multiply spinors by complex numbers, add spinors when they have the same index structure, take outer products of spinors, and contract over spinor indices (one raised and one lowered, both primed or both unprimed.) Since $\epsilon_{AB} \neq 0$, there is a unique spinor ϵ^{AB} which is skew and satisfies

$$\epsilon_{AM}\epsilon^{BM} = \delta^B{}_A \tag{290}$$

where $\delta^B{}_A$ is the unit spinor (defined by $\delta^B{}_A\xi^A = \xi^B$ for all ξ^A.) We can now "raise and lower spinor indices" (i.e., define isomorphisms between $\mathfrak{C}$ and $\mathfrak{C}^*$ and between $\bar{\mathfrak{C}}$ and $\bar{\mathfrak{C}}^*$):

$$\begin{aligned} \xi^A &= \epsilon^{AB}\xi_B & \qquad \xi_A &= \xi^B\epsilon_{BA} \\ \eta^{A'} &= \bar{\epsilon}^{A'B'}\eta_{B'} & \qquad \eta_{A'} &= \eta^{B'}\bar{\epsilon}_{B'A'} \end{aligned} \tag{291}$$

(Note the placement of indices in (291).) Similarly, one can raise and lower an index of a spinor with more than one index. Note that, since ϵ_{AB} is skew, we have

$$\xi^A\sigma_A = -\xi_A\sigma^A \tag{292}$$

Let V denote the collection of all spinors $\xi^{AA'}$ which are real:

$$\overline{\xi^{AA'}} = \xi^{AA'} \tag{293}$$

We can certainly add two elements of V obtain another element of V, and multiply an element of V by a real number to obtain another element of V. (Note that multiplication by a complex number does not preserve (293): $\overline{(\alpha\xi^{AA'})} = \bar{\alpha}\overline{\xi^{AA'}}$.) Thus, V is a real vector space, which, as can easily be checked by introducing a basis, is four-dimensional. (The collection of all Hermitian 2×2 matrices is a real four-dimensional vector space.) We can furthermore define on V an inner product (i.e., a metric) via

$$\xi^{AA'}\xi_{AA'} \tag{294}$$

By introducing a basis, or otherwise, it is easily checked that the signature of this metric is $(+,-,-,-)$.

So far, spinor space is a purely mathematical construct. In order to actually use the spinor space in physics, we must somehow tie it down to our space-time — Minkowski space. This is accomplished, of course, through the vector space V of solutions of (293). The vectors at a point x of Minkowski space form a four-dimensional vector space on which there is a metric of signature $(+,-,-,-)$. We "tie down" spinor space, therefore, by specifying some metric-preserving isomorphism between V and this vector space of vectors at the point x. We assume that such an isomorphism has been fixed once and for all. Thus, we can regard a tensor in Minkowski space at x, e.g., $T^{ab}{}_c$, defining a spinor, $T^{AA'BB'}{}_{CC'}$ which is real:

$$\overline{T^{AA'BB'}{}_{CC'}} = T^{AA'BB'}{}_{CC'} \tag{295}$$

We shall allow ourselves to write such "equivalent" quantities as equal:

$$T^{ab}{}_{c} = T^{AA'BB'}{}_{CC'} \tag{296}$$

In other words, we are free to replace a lowercase Latin index (tensor index in Minkowski space) by the corresponding uppercase Latin index written twice, once unprimed and once primed. For example, the metric in Minkowski space takes the form (see (294))

$$\eta_{ab} = \epsilon_{AB}\bar{\epsilon}_{A'B'} \tag{297}$$

We may thus regard tensors at x as merely a special case of spinors (those having an equal number of primed and unprimed indices, and which, for a real tensor, are real). Finally, note that translation in Minkowski space defines a metric-preserving isomorphism between the vectors at x and the vectors at any other point y. Hence, we have automatically spinors at the point y. More generally, we have the notion of a spinor field, a spinor function of position in Minkowski space. Tensor fields in Minkowski space are, of course, a special case. We can multiply spinor fields by real or complex scalar fields, and spinor fields (when they have the same index structure), take outer products of spinor fields, and contract appropriate spinor indices.

It is possible, in addition, to take derivatives of spinor fields. Let, for example, $T^{A}{}_{BC'}$ be a spinor field. Let r^m be a vector at x, and let x' be the point whose position vector relative to x is ϵr^m. We define $\nabla_m T^{A}{}_{BC'}$ by

$$r^m \nabla_m T^{A}{}_{BC'} = \lim_{\epsilon \to 0} \frac{T^{A}{}_{BC'}(x') - T^{A}{}_{BC'}(x)}{\epsilon} \tag{298}$$

The replacement of Minkowski tensor indices by spinor indices can, of course, be extended to the index of the derivative operator. That is,

$$\nabla_m T^{A}{}_{BC'} = \nabla_{MM'} T^{A}{}_{BC'} \tag{299}$$

In short, the mechanics of calculating with spinor fields is in no essential way different from that of tensor fields. The one point one has to be careful about is the index locations in contractions (see (292).)

One further question must be discussed. To what extent does the imposition of the notion of spinor fields on Minkowski space enrich the structure of Minkowski space? That is, are there essentially inequivalent spinor structures on Minkowski space? To obtain evidence on this question, consider the collection of vector fields on Minkowski space of the form

$$\xi^A \bar{\xi}^{A'} \tag{300}$$

where $\xi^A(x)$ is a spinor field. This vector field is certainly real, and, from (294) is null. (From (292), $\xi^A \xi_A = 0$.) Furthermore, the inner product of two such fields,

$$(\xi^A \bar{\xi}^{A'})(\eta_A \bar{\eta}_{A'}) = (\xi^A \eta_A)(\bar{\xi}^{A'} \bar{\eta}_{A'}) = (\xi^A \eta_A)\overline{(\xi^B \eta_B)} \tag{301}$$

is necessarily non-negative. Thus, the spinor structure on Minkowski space determines a particular time orientation, which we may specify as being the

"future." (Past-directed null vectors then have the form $-\xi^A\bar{\xi}^{A'}$.) Furthermore, the tensor field on Minkowski space defined by the right side of

$$\epsilon_{abcd} = -i(\epsilon_{AB}\epsilon_{CD}\bar{\epsilon}_{A'C'}\bar{\epsilon}_{B'D'} - \bar{\epsilon}_{A'B}\bar{\epsilon}_{C'D'}\epsilon_{AC}\epsilon_{BD}) \tag{302}$$

is real, totally antisymmetric, and satisfies $\epsilon^{abcd}\epsilon_{abcd} = -24$. Hence, this must be all alternating tensor on Minkowski space. Thus, a spinor structure on Minkowski space induces both temporal and spatial parities on Minkowski space. In fact, this is all the structure induced on Minkowski space by a spinor structure. More precisely, given two metric-preserving isomorphisms between V and vectors in Minkowski space, such that these induce the same spatial and temporal parities on Minkowski space, these are related by a linear mapping from $\mathfrak{C}$ onto $\mathfrak{C}$ which preserves ϵ_{AB} (i.e., by an element of $\mathrm{SL}(2,\mathbb{C})$.) Finally, note that there are precisely two ϵ_{AB}-preserving linear mappings on $\mathfrak{C}$ which leave V (and hence tensors in Minkowski space) invariant, namely the identity and minus the identity. This is the statement of the "two-valuedness" associated with spinors.

Finally, we briefly indicate how one translates formulae from the two-component to the four-component spinor notation. A four-component spinor is a pair of two-component spinors, $(\xi^A, \eta_{A'})$, consisting of one spinor with a primed and one with an unprimed index. One then normally chooses a basis for $\mathfrak{C}$ and writes this pair out as a 4×1 column matrix. The "γ-matrices" in the four-component notation serve the function of combining these components in the appropriate way to obtain the various scalar, vector, and tensor fields on Minkowski space associated with the pair $(\xi^A, \eta_{A'})$. For example, a pair $(\xi^A, \eta_{A'})$ defines the following fields on Minkowski space:

$$\begin{gathered}\xi^A\bar{\eta}_A, \quad \xi_A\bar{\xi}_{A'}, \quad \eta_{A'}\bar{\eta}_A, \quad \xi_A\eta_{A'}, \quad \xi_A\xi_B\bar{\epsilon}_{A'B'}, \\ \eta_{A'}\eta_{B'}\epsilon_{AB}, \quad \xi_A\bar{\eta}_B\bar{\epsilon}_{A'B'}, \quad \xi_B\bar{\xi}_{A'}\epsilon_{AC}\bar{\epsilon}_{C'B'}\end{gathered} \tag{303}$$

The spinor notation discussed here (which is due to Penrose) essentially avoids the γ-matrices by choosing a basis for neither spinor space nor Minkowski space. Questions of (restricted) Lorentz invariance simply do not arise: one cannot, with this notation, write anything which is not invariant.

19 The Dirac Equation

The field which describes a free, massive, spin-$\frac{1}{2}$ particle consists of a pair, $(\xi^A, \eta_{A'})$, of spinor fields on Minkowski space. These fields must satisfy the *Dirac equation*:

$$\nabla_{AA'}\xi^A = \frac{\mu}{\sqrt{2}}\eta_{A'} \tag{304}$$

$$\nabla^{AA'}\eta_{A'} = -\frac{\mu}{\sqrt{2}}\mu\xi^A \tag{305}$$

where μ is a positive real number (which, as we shall see shortly, is essentially the mass of the particle.) The Dirac equations are to be considered as analogous to the Klein-Gordon equation, or to the Maxwell equations.

There is another, for some purposes more illuminating, form for Dirac's equations. Taking the derivative of (304), we have

$$\nabla^{BA'}\nabla_{AA'}\xi^A = \frac{\mu}{\sqrt{2}}\nabla^{BA'}\eta_{A'} \tag{306}$$

Substituting (305) on the right in (306), and using the fact that

$$\nabla^{BA'}\nabla_{AA'} = \frac{1}{2}\delta^B{}_A\Box \tag{307}$$

we obtain

$$(\Box + \mu^2)\xi^B = 0 \tag{308}$$

Thus, the Dirac equations imply that the spinor field ξ^A (and, by a similar argument, $\eta_{A'}$) satisfies the Klein-Gordon equation. Conversely, if ξ^A is any solution of (308), then, defining $\eta_{A'}$ by (304) (note $\mu \neq 0$), $(\xi^A, \eta_{A'})$ is a solution of Dirac's equations. In other words, there is a one-to-one correspondence between solutions of Dirac's equations and solutions of (308). Why, then, do we choose to deal with a pair of spinor fields and the relatively complicated equations (304), (305) rather than simply a simple spinor field and (308)? The reason is that there is a certain symmetry between ξ^A and $\eta_{A'}$ which, while merely a curiosity at present, will later be found to be related to the discrete symmetries of Minkowski space.

One further consequence of (308) is that it makes clear the fact that the problem of finding solutions of Dirac's equations is no more and no less difficult than that of finding solutions of the Klein-Gordon equations. Fix two constant spinor fields, α^A and β^A, on Minkowski space. Then, by the remarks above, each solution, $(\xi^A, \eta_{A'})$, of Dirac's equations defines two solutions, $\xi^A\alpha_A$ and $\xi^A\beta_A$, of the Klein-Gordon equation, and, conversely, if ϕ and ψ are solutions of the Klein-Gordon equation, then $\phi\alpha^A + \psi\beta^A$ is a solution of (308), and hence defines a solution of Dirac's equations.

Note that if $(\xi^A, \eta_{A'})$ is a solution of Dirac's equations, so is $(\bar{\eta}^A, \bar{\xi}_{A'})$. We call $(\bar{\eta}^A, \bar{\xi}_{A'})$ the *complex-conjugate* of the solution $(\xi^A, \eta_{A'})$ (analogous to the complex-conjugate of a solution of the Klein-Gordon equation). Of course, complex-conjugation, applied twice to a solution of Dirac's equations, yields the original solution.

We now go to momentum space. Set

$$\xi^A(x) = \int_{M_\mu} \xi^A(k)e^{ik_b x^b}\,\mathrm{d}V_\mu \tag{309}$$

$$\eta_{A'}(x) = \int_{M_\mu} \eta_{A'}(k)e^{ik_b x^b}\,\mathrm{d}V_\mu \tag{310}$$

where $\xi^A(k)$ and $\eta_{A'}(k)$ are spinor-valued functions. These functions are only defined on M_μ, and the integrals (309), (310) are only carried out over M_μ, because of (308). Inserting (309) and (310) into (304) and (305), we obtain

$$ik_{AA'}\xi^A(k) = \frac{\mu}{\sqrt{2}}\eta_{A'}(k) \tag{311}$$

$$ik^{AA'}\eta_A(k) = -\frac{\mu}{\sqrt{2}}\xi^A(k) \tag{312}$$

Note that each of (311) and (312) implies the other. Thus, a solution of Dirac's equations is characterized by a pair of spinor-valued functions, $\xi^A(k)$ and $\eta_{A'}(k)$, on M_μ, subject to (311) and (312). (Alternatively, a solution is characterized by a single, arbitrary, spinor function $\xi^A(k)$ on M_μ. Then $\eta_{A'}(k)$ is defined by (311), and (312) follows identically.) A solution of Dirac's equations is said to be *positive-frequency* (resp. *negative-frequency*) if $\xi^A(k)$ and $\eta_{A'}(k)$ vanish on M_μ^- (resp. M_μ^+). In momentum space, complex conjugation has the effect

$$\xi^A(k) \to \bar{\eta}^A(-k) \qquad\qquad \eta_{A'}(k) \to \bar{\xi}_{A'}(-k) \tag{313}$$

Thus, just as in the Klein-Gordon case, complex-conjugation takes positive-frequency solutions to negative-frequency solutions, and vice-versa. (Roughly speaking, positive-frequency solutions represent electrons, and negative-frequency solutions positrons.)

Let $(\xi^A, \eta_{A'})$ be a solution of Dirac's equations, and consider the real vector field

$$j^a = \xi^A\bar{\xi}^{A'} + \eta^{A'}\bar{\eta}^A \tag{314}$$

in Minkowski space. First note that, since each term on the right in (314) is a future-directed null vector, j^a is future-directed and either timelike or null. We have, for the divergence of j^a,

$$\nabla_a j^a = \xi^A\nabla_{AA'}\bar{\xi}^{A'} + \bar{\xi}^{A'}\nabla_{AA'}\xi^A + \eta^{A'}\nabla_{AA'}\bar{\eta}^A + \bar{\eta}^A\nabla_{AA'}\eta^{A'} \tag{315}$$

Substituting (304) and (305), and using (292), we find

$$\nabla_a j^a = 0 \tag{316}$$

Thus, j^a is a real, future-directed timelike or null, divergence-free vector field. Therefore, the integral of j^a over a spacelike 3-plane yields a nonnegative number which, assuming that the Dirac field goes to zero sufficiently quickly at infinity, is independent of the choice of the 3-plane. This integral can be used to define a norm on solutions of Dirac's equations. The situation is much simpler when translated into momentum space (see (23)). We define the norm by

$$\frac{i\sqrt{2}}{\mu}\left(\int_{M_\mu^+}\xi^A(k)\bar{\eta}_A(k)\,\mathrm{d}V_\mu - \int_{M_\mu^-}\xi^A(k)\bar{\eta}_A(k)\,\mathrm{d}V_\mu\right) \tag{317}$$

Note that, because of (311) and (312), the expression (317) is equal to both of

$$\left(\int\limits_{M_\mu^+} - \int\limits_{M_\mu^-} \right) \frac{1}{\mu^2} (\xi^A \bar{\xi}^{A'} + \bar{\eta}^A \eta^{A'}) k_{AA'} \, \mathrm{d}V_\mu \tag{318}$$

$$\left(\int\limits_{M_\mu^+} - \int\limits_{M_\mu^-} \right) \frac{2}{\mu^2} \xi^A \bar{\xi}^{A'} k_{AA'} \, \mathrm{d}V_\mu \tag{319}$$

The forms (318) or (319) show, in particular, that our norm is positive. (The vector $\xi^A \bar{\xi}^{A'}$ is future-directed null, whereas $k_{AA'}$ is future-directed timelike on M_μ^+ and past-directed timelike on M_μ^-.)

We have now obtained a norm on our collection of solutions of Dirac's equation. In order to obtain a Hilbert space, therefore, we have only to impose the structure of a complex vector space on our collection of solutions. In other words, we must define addition of solutions and multiplication of solutions by complex numbers. There is only one reasonable way to define addition: one simply adds the corresponding spinor fields in Minkowski space (or, in momentum space, adds the corresponding spinor functions on the mass shell.) One might think, at first glance, that there is also only one reasonable definition of the product of a complex number and a solution of Dirac's equations: if α is a complex number, and $(\xi^A(k), \eta_{A'}(k))$ is a solution of Dirac's equations, one defines the product to be the solution $(\alpha\xi^A(k), \alpha\eta_{A'}(k))$. In other words, since the Dirac equation is linear on (complex) spinor fields, the solutions of this equation naturally have the structure of a complex vector space. There is, however, an alternative way to define the product of a solution of Dirac's equations and a complex number. Let $\xi(k)$ and $\eta_{A'}(k)$ be a pair of spinor functions on M_μ which satisfy (311) and (312), i.e., a solution (in momentum space) of Dirac's equations. Let α be a complex number. Then we might also define the product of α and $(\xi^A(k), \eta_{A'}(k))$ to be the solution

$$\begin{array}{ll} (\alpha\xi^A(k), \alpha\eta_{A'}(k)) & \text{for } k \in M_\mu^+ \\ (\bar{\alpha}\xi^A(k), \bar{\alpha}\eta_{A'}(k)) & \text{for } k \in M_\mu^- \end{array} \tag{320}$$

of Dirac's equations. That is to say, we multiply the positive-frequency part of the fields by α and the negative-frequency part by $\bar{\alpha}$. We obtain, in this way, an essentially different complex vector space of solutions of Dirac's equations. In fact, we adopt this second — rather less aesthetic — alternative. As we shall see later, this choice is essential to obtain agreement between theory and experiment.

We now have a complex vector space with a norm, (317), and hence a Hilbert space. More precisely, the Hilbert space of the Dirac equation, H_D, is the collection of all pairs, $(\xi^A(k), \eta_{A'}(k))$, of spinor functions on M_μ which satisfy (311), which are measurable, and for which the integral (317) converges. The

inner product on our Hilbert space can now be obtained from the norm via the identity

$$(\phi,\psi)=\frac{1}{4}\left(\|\phi+\psi\|^2-\|\phi-\psi\|^2\right)+\frac{i}{4}\left(\|\psi+i\psi\|^2-\|\phi-i\psi\|^2\right) \tag{321}$$

Using (317) and (321), the inner product on H_D takes the form

$$\frac{1}{\mu^2}\int\limits_{M_\mu^+}\left(\xi^A\bar{\sigma}^{A'}+\eta^{A'}\bar{\tau}^A\right)k_{AA'}\,\mathrm{d}V_\mu-\frac{1}{\mu^2}\int\limits_{M_\mu^-}\left(\bar{\xi}^{A'}\sigma^A+\bar{\eta}^A\tau^{A'}\right)k_{AA'}\,\mathrm{d}V_\mu \tag{322}$$

where $(\xi^A(k),\eta_{A'}(k))$ and $(\sigma^A(k),\tau_{A'}(k))$ are two solutions of Dirac's equations. Note the appearance of the complex-conjugations in the integral over M_μ^-. These arise because of our choice of the complex vector space structure for H_D.

To summarize, whereas the solutions of Dirac's equations have only one reasonable real vector space structure and only one reasonable norm, there are two possible complex vector space structures, of which we choose one. This choice then leads to the particular form for the inner product on our Hilbert space.

We now introduce the antisymmetric Fock space based on H_D. We thus have creation and annihilation operators, number operators, etc.

In the real Klein-Gordon and Maxwell cases, we were dealing with real fields on Minkowski space. This feature was reflected in momentum space by our requirement that the fields on the mass shell be invariant under simultaneous complex-conjugation and reflection through the origin. Physically, we were dealing with particles which are identical with their antiparticles. While we could, of course, restrict ourselves to real ($\xi^A=\bar{\eta}^A$) solutions of Dirac's equations, it is convenient not to do so. Thus, the functions on the future mass shell need bear no special relation to those on the past mass shell. This state of affairs leads to a pair of projection operators on H_D. Let $(\xi^A(k),\eta_{A'}(k))\in H_D$. Then the action of P^+ (projection onto the positive frequency part) is defined by

$$P^+(\xi^A(k),\eta_{A'}(k))=\begin{cases}(\xi^A,\eta_{A'}) & \text{for}\quad k\in M_\mu^+\\ (0,0) & \text{for}\quad k\in M_\mu^-\end{cases} \tag{323}$$

and similarly for P^-. Note that

$$P^++P^-=\mathbb{I} \tag{324}$$

These operators are both projection operators, i.e., they are defined everywhere and satisfy

$$(P^+)^2=P^+\qquad(P^-)^2=P^- \tag{325}$$

Eigenstates of P^+ with eigenvalue one (i.e., positive-frequency solutions) are called *particle states*, and those of P^- *antiparticle states*. Thus, we can speak of creation or annihilation of particle and antiparticle states, number operators

for particles and antiparticles, etc. When we discuss charge, for example, we shall introduce the total charge operator, $eP^- - eP^+$, (this is the form when the particles have negative charge, e.g., electrons) where e is the fundamental charge.

The Dirac equation describes particles of mass $\hbar\mu$ and spin $\frac{1}{2}$. This statement must, of course, be proven using the techniques of Sect. 17. We now give the proof. The only piece of additional machinery we require is the notion of the Lie derivative of a spinor field. Quite generally, any smooth mapping, with smooth inverse, from Minkowski space to itself takes any tensor field on Minkowski space to another tensor field on Minkowski space. Smooth mappings which "differ infinitesimally from the identity mapping" are described by smooth vector fields. The corresponding "infinitesimal change in a tensor field" defines the Lie derivative of that tensor field. Does a smooth mapping, with smooth inverse, on Minkowski space take spinor fields to spinor fields? In other words, can we formulate a natural notion of the Lie derivative of a spinor field (by a vector field) so that the Lie derivative of a tensor field will arise as a special case (i.e., considering a tensor field as merely a special case of a spinor field when the numbers of primed and unprimed spinor indices are equal)? Unfortunately, the answer to these questions is no. To see this, suppose for a moment that it were possible to generalize the Lie derivative from tensor to spinor fields. Let v^a be an arbitrary smooth vector field on Minkowski space. Then we would have

$$\mathcal{L}_v \eta_{ab} = \mathcal{L}_v(\epsilon_{AB}\bar{\epsilon}_{A'B'}) = \epsilon_{AB}\mathcal{L}_v\bar{\epsilon}_{A'B'} + \bar{\epsilon}_{A'B'}\mathcal{L}_v\epsilon_{AB} \tag{326}$$

But, since ϵ_{AB} is skew, so must be $\mathcal{L}_v\epsilon_{AB}$, and similarly for $\mathcal{L}_v\bar{\epsilon}_{A'B'}$. Thus, the right side of (326) must be some multiple of the Minkowski metric η_{ab}. But it is simply false that, for an arbitrary smooth vector field v^a on Minkowski space, $\mathcal{L}_v\eta_{ab}$ is a multiple of η_{ab}. Thus, we cannot in general define the Lie derivative of a spinor field. Intuitively, the problem is that the light-cone structure of Minkowski space is an essential ingredient in the very definition of a spinor field. A smooth (finite or infinitesimal) mapping on Minkowski space which alters the light-cone structure simply does not know what to do with a general spinor field.

The remarks above are also the key to resolving the problem. In order to define spin and mass, it is only necessary to take Lie derivatives of spinor fields by vector fields v^a which satisfy (216) — i.e., by vector fields which do preserve the light-cone structure of Minkowski space. We might expect to be able to define Lie derivatives by such vector fields, and this is indeed the case. The formula is, for example,

$$\begin{aligned}\mathcal{L}_v T^{ABC'}{}_{DE'} &= v^m\nabla_m T^{ABC'}{}_{DE'} - \frac{1}{2}T^{MBC'}{}_{DE'}\nabla_{MM'}v^{AM'} \\ &\quad - \frac{1}{2}T^{AMC'}{}_{DE'}\nabla_{MM'}v^{BM'} - \frac{1}{2}T^{ABM'}{}_{DE'}\nabla_{MM'}v^{MC} \\ &\quad + \frac{1}{2}T^{ABC'}{}_{ME'}\nabla_{DM'}v^{MM'} + \frac{1}{2}T^{ABC'}{}_{DM'}\nabla_{ME'}v^{MM'}\end{aligned} \tag{327}$$

Note that Lie differentiation commutes with complex-conjugation (v^a is real), raising and lowering of spinor indices, and contraction of spinor indices. Note, furthermore, that (327) reduces to the usual Lie derivative for tensor fields. It follows from the remarks on p. 65 that (327) is the only formula which satisfies these properties.

We first determine the mass associated with the Dirac equation. Let $(\xi^A(x), \eta_{A'}(x))$ be a solution of the Dirac equation, and r^a a vector at some point x of Minkowski space. Then

$$\begin{aligned}(r^a \underline{P}_a)\xi^M &= \frac{\hbar}{i} r^a \nabla_a \xi^M \\ (r^a \underline{P}_a)(r^b \underline{P}_b)\xi^M &= \hbar^2 r^a r^b \nabla_a \nabla_b \xi^M \end{aligned} \tag{328}$$

Substituting η^{ab} for $r^a r^b$, we obtain

$$\underline{P}_a \underline{P}^a \xi^M = -\hbar^2 \Box \xi^M = \hbar^2 \mu^2 \xi^M \tag{329}$$

where we have used (308). Hence, from (268), the mass associated with the Dirac equation is $\hbar\mu$. The spin calculation is slightly more complicated. Let s^{cd} be a skew tensor at x. Then, from (327),

$$s^{cd} \underline{P}_{cd} \xi^M = \frac{\hbar}{i} \left(s^c{}_d x^d \nabla_c \xi^M - \frac{1}{2} \xi^N s^{MN'}{}_{NN'} \right) \tag{330}$$

If r^b is a vector at x, we have, therefore.

$$\begin{aligned} r^b \underline{P}_b s^{cd} \underline{P}_{cd} \xi^M &= \hbar^2 r^b \nabla_b \left(s^c{}_d x^d \nabla_c \xi^M - \frac{1}{2} \xi^N s^{MN'}{}_{NN'} \right) \\ &= \hbar^2 r^{BB'} s^{CC'DD'} \left(\epsilon_{BD} \bar{\epsilon}_{B'D'} \nabla_{CC'} \xi^M + \frac{1}{2} \bar{\epsilon}_{D'C'} \delta^M{}_C \nabla_{BB'} \xi_D \right) \end{aligned} \tag{331}$$

Substituting $u_a \epsilon^a bcd$ for $r^b s^{cd}$, and using (269) and (302),

$$\begin{aligned} u^a \underline{W}_a \xi^M &= i\hbar^2 \left(\epsilon^{AB} \epsilon^{CD} \bar{\epsilon}^{A'C'} \bar{\epsilon}^{B'D'} - \bar{\epsilon}^{A'B'} \bar{\epsilon}^{C'D'} \epsilon^{AC} \epsilon^{BD} \right) u_{AA'} \\ &\qquad \times \left(\epsilon_{BD} \bar{\epsilon}_{B'D'} \nabla_{CC'} \xi^M + \frac{1}{2} \bar{\epsilon}_{D'C'} \delta^M{}_C \nabla_{BB'} \xi_D \right) \\ &= -\frac{i}{2} \hbar^2 u^a \nabla_a \xi^M + i\hbar^2 u^{MA'} \nabla_{AA'} \xi^A \end{aligned} \tag{332}$$

Therefore,

$$\begin{aligned} u^a \underline{W}_a u^b \underline{W}_b \xi^M = (i\hbar^2)^2 \Big(&\frac{1}{4} u^a u^b \nabla_a \nabla_b \xi^M - \frac{1}{2} u^{MA'} u^b \nabla_{AA'} \nabla_b \xi^A \\ &- \frac{1}{2} u^{MB'} u^a \nabla_{BB'} \nabla_a \xi^B + u^{MA'} u^{AB'} \nabla_{AA'} \nabla_{BB'} \xi^B \Big) \end{aligned} \tag{333}$$

Finally, substituting η^{ab} for $u^a u^b$, we have

$$\underline{W}^a \underline{W}_a \xi^M = \frac{3}{4} \hbar^4 \Box \xi^M = -\frac{3}{4} \hbar^2 m^2 \xi^M \tag{334}$$

We conclude from (270) that the Dirac equation describes a particle with spin $\frac{1}{2}$.

20 The Neutrino Equation

A neutrino is essentially a "massless Dirac particle". There are, however, a few features which are particular to the case $\mu = 0$.

The (four-component) neutrino field consists of a pair, $(\xi^A, \eta_{A'})$ of spinor fields on Minkowski space, subject to the *neutrino equation* (see (304), (305)):

$$\nabla_{AA'}\xi^A = 0 \tag{335}$$

$$\nabla^{AA'}\eta_{A'} = 0 \tag{336}$$

Note that, whereas in the massive case either of the two spinor fields can be obtained from the other (via (304), (305)), the fields become "uncoupled" in the massless case. That is to say, each spinor field satisfies its own equation. Taking a derivative of (335),

$$\nabla^{BA'}\nabla_{AA'}\xi^A = 0 \tag{337}$$

and using (307), we obtain

$$\Box\xi^A = 0 \tag{338}$$

and similarly for $\eta_{A'}$. Thus, each of our neutrino fields satisfies the wave equation. Note, however, that (338) does not imply (335). (Solutions of the neutrino equations can, however, be obtained from solutions of the wave equation. If $\alpha_{A'}$ satisfies the wave equation, then $\xi^A = \nabla^{AA'}\alpha_{A'}$ satisfies (335).)

The *complex-conjugate* of the solution $(\xi^A, \eta_{A'})$ of the neutrino equation is the solution $(\bar{\eta}^A, \bar{\xi}_{A'})$.

Passing to momentum space, we set

$$\xi^A(x) = \int_{M_0} \xi^A(k) e^{ik_b x^b}\, \mathrm{d}V_0 \tag{339}$$

$$\eta_{A'}(x) = \int_{M_0} \eta_{A'}(k) e^{ik_b x^b}\, \mathrm{d}V_0 \tag{340}$$

where $\xi^A(k)$ and $\eta_{A'}(k)$ are spinor-valued functions on the zero-mass shell, M_0. In momentum space, (335) and (336) become

$$\xi^A(k)k_{AA'} = 0 \tag{341}$$

$$\eta_{A'}(k)k^{AA'} = 0 \tag{342}$$

Positive-frequency and negative-frequency solutions of the neutrino equations are well-defined. Complex-conjugation again reverses frequency, and is again expressed in momentum apace by the equations (313).

The current (314) is still divergence-free in the massless case. (In fact, the proof is rather simpler with $\mu = 0$.) This fact leads to a norm on solutions of the neutrino equation. The simplest way to obtain the norm, however, is as a

"$\mu \to 0$ limit" of the Dirac norm. Consider (318). It is not difficult to check from (311) and (312) that

$$\xi^A(k)\bar{\xi}^{A'}(k) + \eta^{A'}\bar{\eta}^A(k) = \alpha(k)k^{AA'} \tag{343}$$

where $\alpha(k)$ is a real function on M_μ which is positive on M_μ^+ and negative on M_μ^-. From (318), the norm in the Dirac case is simply

$$\int_{M_0} |\alpha(k)| \, \mathrm{d}V_\mu \tag{344}$$

We now return to the massless case. Eqn. (341) implies that $\xi^A(k)\bar{\xi}^{A'}(k)$ is proportional to $k^{AA'}$, while (342) implies that $\eta^{A'}(k)\bar{\eta}^A(k)$ is also proportional to $k_{AA'}$. Therefore,

$$\xi^A(k)\bar{\xi}^{A'}(k) + \eta^{A'}\bar{\eta}^A(k) = \alpha(k)k^{AA'} \tag{345}$$

on M_0, where $\alpha(k)$ is real on M_0 and positive on M_0^+ and negative on M_0^-. We therefore define the norm in the neutrino case, in analogy with (344), by

$$\int_{M_0} |\alpha(k)| \, \mathrm{d}V_\mu \tag{346}$$

For the complex vector space structure in the massless case, we use the same convention as in the massive case (see (320)).

In fact, the theory we have been discussing is not very interesting physically. The reason is that our Hilbert space of solutions of the neutrino equation contains four irreducible subspaces: positive-frequency solutions with $\eta_{A'} = 0$, negative-frequency solutions with $\eta_{A'} = 0$, positive-frequency solutions with $\xi^A = 0$, and negative-frequency solutions with $\xi^A = 0$. Every solution can be written uniquely as the sum of four solutions, one from each class above. Thus, our neutrino field describes four similar particles. But neutrinos in the real world appear in pairs (particle-antiparticle.) Thus, we would like to introduce a field whose Hilbert space has only two irreducible (under the restricted Poincaré group) subspaces. The result is what is called the "two-component neutrino theory", which we now describe. (The only purpose in treating the four-component theory at all was to make explicit the analogy with the Dirac equation.)

The (two-component) neutrino field is a single spinor field ξ^A on Minkowski space which satisfies (335), and, therefore, (338). In momentum space, we have a spinor-valued function $\xi^A(k)$ on M_0 which satisfies (341). This equation implies

$$\xi^A(k)\bar{\xi}^{A'}(k) = \alpha(k)k^{AA'} \tag{347}$$

where $\alpha(k)$ is real, and positive on M_0^+ and negative on M_0^-. We define the norm on our solutions by

$$\int_{M_0} |\alpha(k)| \, \mathrm{d}V_0 \tag{348}$$

The complex vector space structure is defined as before: the product of a complex number β and a solution $\xi^A(k)$ is defined to be the solution

$$\begin{array}{ll} \beta\xi^A(k) & \text{for } k \in M_0^+ \\ \bar{\beta}\xi^A(k) & \text{for } k \in M_0^- \end{array} \tag{349}$$

The collection of all measurable spinor functions $\xi^A(k)$ on M_0 for which (341) is satisfied and (348) converges, with the above complex vector space structure, is a Hilbert space which we write as H_N.

We introduce the antisymmetric Fock space based on H_N. We thus have creation and annihilation operators, number operators, etc.

We introduce on H_N the two projection operators P^+ and P^-, projection onto positive and negative frequency, respectively. These operators, of course, satisfy (324) and (325).

Finally, we remark on the spin and mass to be associated with H_N. We have done the mass calculation several times: (338) clearly leads to $m = 0$ for H_N. Furthermore, most of the work involved in calculating the spin has already been done. Nowhere in the argument leading to (332) did we use the fact that ξ^A satisfies the Dirac equation, and so (332) holds also in the neutrino case. But now (335) implies that the second term on the right in (332) vanishes, so we have

$$u^a \underline{W}_a \xi^B = \frac{1}{2}\frac{\hbar^2}{i} u^a \nabla_a \xi^B \tag{350}$$

Since, furthermore,

$$u^a \underline{P}_a \xi^B = \frac{\hbar}{i} u^a \nabla_a \xi^B \tag{351}$$

one is tempted to conclude from (271) that $s = \frac{1}{2}$ for H_N. This conclusion is essentially correct, but one technical point must be clarified. (Unfortunately, our notation is rather badly suited to the remarks below, and so they will sound rather mystical.) The problem involves what the i's mean in (350) and (351). (This problem never arose in the Dirac case because the i's were always squared away, so their meaning was irrelevant.) Where did the i's come from? The i in (351) came from the $\hbar/i$ factors which are introduced in the operator fields associated with the Poincaré group. This i means "multiplication by i within the Hilbert space H_N" because only in this way does one obtain Hermitian operators from "infinitesimal unitary operators". In other words, the i in (351) arises from very general considerations involving the action of a Lie group on a Hilbert space, and, in this general framework within which the formalism was set up, there is only one notion of multiplication by i, namely, multiplication within the Hilbert space. Thus, the "i" in (351) multiplies the positive-frequency part of what follows by i, and the negative-frequency part by $-i$. (See (349).) The i in (350), on the other hand, is a quite different animal. It arose from the i in (302). (The i's in $\underline{P}_a$ and $\underline{P}_{ab}$ (see (269)) combine to give -1.) But the i in (302) appears because of the way that the real tensor field ϵ^{abcd} must be expressed in terms of spinors. Hence, the "i" in (350), because of its origin, represents simply multiplication of a tensor field by i. That is to say, the "i-operators" in

(350) and (351) are equal for positive-frequency solutions, and minus each other for negative-frequency solutions. Thus, $s = \frac{1}{2}$ for positive-frequency solutions (neutrinos), and $s = -\frac{1}{2}$ for negative-frequency solutions (antineutrinos). That is, in the neutrino case the particle and its antiparticle have opposite helicity. This "prediction" is in fact confirmed by experiment.

21 Complex Klein-Gordon Fields

In Sect. 5, we dealt with real equations of the Klein-Gordon equation (although, for reasons of motivation, we chose to characterize such fields as complex positive-frequency solutions). Such fields describe particles with spin zero which are identical with their antiparticles (e.g., the π^0). On the other hand, there are spin-zero particles which are not identical with their antiparticles (the π^+ and π^-). These are described by complex solutions of the Klein-Gordon equation.

Consider a complex scalar field, $\phi(x)$, in Minkowski space which satisfies the Klein-Gordon equation, (5). In momentum space,

$$\phi(x) = \int\limits_{M_\mu} \phi(k) e^{ik_b x^b} \, \mathrm{d}V_\mu \tag{352}$$

Thus, our solution is characterized by a complex-valued function $\phi(k)$ on M_μ. (In the real case, one requires in addition $\phi(-k) = \bar{\phi}(k)$.) The norm of such a function is defined by

$$\frac{1}{\hbar} \int\limits_{M_\mu} \phi(k) \bar{\phi}(k) \, \mathrm{d}V_\mu \tag{353}$$

We adopt, for the complex vector space structure on these functions, essentially the same structure used in the Dirac and neutrino case. To "multiply" $\phi(k)$ by a complex number, one takes

$$\begin{array}{ll} \alpha\phi(k) & \text{for } k \in M_\mu^+ \\ \bar{\alpha}\phi(k) & \text{for } k \in M_\mu^- \end{array} \tag{354}$$

The collection of all measurable, square-integrable (in the sense of (353), complex-valued functions on M_μ, with this complex vector space structure, is our Hilbert space, H_{CKG}, for complex solutions of the Klein-Gordon equation. (We shall write the Hilbert space of Sect. 5 as H_{RKG}.)

There is defined on H_{CKG} the two projection operators, P^+ and P^-, which take the positive-frequency part and negative-frequency part, respectively.

We introduce the symmetric Fock space based on H_{CKG}, creation and annihilation operators, number operators, etc.

22 Positive Energy

Many of the quantities associated with an elementary particle (e.g., charge) are reversed in the passage from a particle to its antiparticle. It is observed

experimentally, however, that energy is not one of these quantities. For example, if an electron and a positron annihilate (say, with negligible kinetic energy), then the total, energy released is $2m$, and not zero. We are thus forced to assign a (rest) energy $+m$ to both a positron and an electron. Where does this fact appear in our formalism?

Of course, "energy" refers to the state of motion of an observer. This "state of motion" is represented by some constant, unit, future-directed timelike vector field r^a in Minkowski space. The energy operator is then $\underline{E} = r^a \underline{P}_a$. It should be emphasized that we are not free to assign energies arbitrarily to obtain agreement with experiment. The very concept of energy is based in an essential way on the action of the Poincaré group (more explicitly, on the time translations). If we wish to avoid a radical change in what energy means in the passage from classical to quantum theory, we must choose for the energy in quantum field theory that quantity which arises naturally from time translations in Minkowski space, i.e., we must choose the $\underline{E}$ above. We take as our precise statement that "energies are nonnegative" the statement that the expectation value of $\underline{E}$ in any state σ (on which $\underline{E}$ is defined) be nonnegative:

$$(\sigma, \underline{E}\sigma) \geq 0 \tag{355}$$

Is it true or false that (355) holds for the five Hilbert spaces we have constructed, H_{RKG}, H_{CKG}, H_M, H_D, H_N?

We begin with the real Klein-Gordon case. The Hilbert space consists of measurable, square-integrable, complex-valued functions $\phi(k)$ on M_μ which satisfy

$$\phi(-k) = \bar{\phi}(k) \tag{356}$$

Such functions do not have an obvious complex vector space structure. If $\phi(k)$ satisfies (356), and α is a complex number, then $\alpha\phi(k)$ will not in general satisfy (356). This fact, of course, is not surprising: there is no obvious way to take the "product" of a complex number and a real solution of a differential equation to obtain another real solution. This problem is resolved, in H_{RKG}, by choosing one of the two mass shells to be preferred, and calling it the "future" mass shell, M_μ^+. We then agree that, to multiply by α, "M_μ^+ gets α while M_μ^- must be content with $\bar{\alpha}$." In other words, we define multiplication of $\phi(k)$ by α by

$$\begin{array}{ll} \alpha\phi(k) & k \in M_\mu^+ \\ \bar{\alpha}\phi(k) & k \in M_\mu^- \end{array} \tag{357}$$

It should be emphasized that, in the real case, we are forced (by the requirement that we obtain a Hilbert space) to select one preferred mass shell and define multiplication by (357).

Now consider the energy. If $\phi(x)$ is a real solution of the Klein-Gordon equation, then

$$\underline{E}\phi(x) = \frac{\hbar}{i} r^a \nabla_a \phi(x) \tag{358}$$

Because of the i in (358), one might naively think that (358) does not represent a real solution of the Klein-Gordon equation, and so that (358) is not a definition

for $\underline{E}$. This, of course, is not the case. The i in (358) arose because of general considerations involving representations of the Poincaré group (Sect. 16), and means "multiplication within the Hilbert space H_{RKG}." In momentum space, the operator $\hbar r^a \nabla_a$ has the effect

$$\phi(k) \to i\hbar(r^a k_a)\phi(k) \tag{359}$$

Note that (359) does not destroy (356), a statement which reflects the fact that

$$\phi(x) \to \hbar r^a \nabla_a \phi \tag{360}$$

is an unambiguous operation on real solutions of the Klein-Gordon equation. Now using (357), the energy operator in momentum space takes the form

$$\phi(k) \to \begin{cases} \hbar r^a k_a \phi(k) & k \in M_\mu^+ \\ -\hbar r^a k_a \phi(k) & k \in M_\mu^- \end{cases} \tag{361}$$

The expectation value of $\underline{E}$ in the state $\phi(k)$ is

$$\int\limits_{M_\mu^+} (r^a k_a)\phi(k)\bar{\phi}(k)\, \mathrm{d}V_\mu - \int\limits_{M_\mu^-} (r^a k_a)\phi(k)\bar{\phi}(k)\, \mathrm{d}V_\mu \tag{362}$$

which, of course, is positive. (Why don't we just define the energy operator by (359), (360), leaving out the i? Because the expectation value of this operator is not real. That is, the i is needed for Hermiticity.)

We summarize the situation. In order to make a Hilbert space of real solutions of the Klein-Gordon equation, we are forced to select a preferred mass shell to be called "future". Then, provided r^a is "future-directed" according to this convention, $\underline{E}$ will have positive expectation values.

Now consider the complex Klein-Gordon case. The energy operator still has the form

$$\underline{E}\phi(x) = \frac{\hbar}{i} r^a \nabla_a \phi(x) \tag{363}$$

and i still means multiplication within our Hilbert space. In momentum space, $\phi(k)$ is an arbitrary measurable, square-integrable, complex-valued function. The operator $\hbar r^a \nabla_a$ has the effect

$$\phi(k) \to i\hbar(r^a k_a)\phi(k) \tag{364}$$

We must still multiply (364) by $1/i$. But we now have the freedom to select one of two possible complex vector space structures on the complex solutions of the Klein-Gordon equation. For the "product" of a complex number α and $\phi(k)$, we could choose

$$\alpha\phi(k) \tag{365}$$

or, alternatively,

$$\begin{matrix} \alpha\phi(k) & k \in M_\mu^+ \\ \bar{\alpha}\phi(k) & k \in M_\mu^- \end{matrix} \tag{366}$$

The resulting energy operators are

$$\underline{E}\phi(k) = \hbar(r^a k_a)\phi(k) \tag{367}$$

$$\underline{E}\phi(k) = \begin{cases} \hbar r^a k_a \phi(k) & k \in M_\mu^+ \\ -\hbar r^a k_a \phi(k) & k \in M_\mu^- \end{cases} \tag{368}$$

respectively. Finally, the resulting expectation values of $\underline{E}$ are

$$\int\limits_{M_\mu} r^a k_a \phi(k)\bar{\phi}(k)\, \mathrm{d}V_\mu \tag{369}$$

$$\int\limits_{M_\mu^+} (r^a k_a)\phi(k)\bar{\phi}(k)\, \mathrm{d}V_\mu - \int\limits_{M_\mu^-} (r^a k_a)\phi(k)\bar{\phi}(k)\, \mathrm{d}V_\mu \tag{370}$$

respectively. But note that (369) can take both positive and negative values, while (370) is always nonnegative. But this is exactly what one might expect. The complex vector space structure (365) does not prefer one time direction over the other: it makes no reference to past and future. Therefore, it could not possibly lead to a positive energy, for the energy associated with r^a is certainly minus the energy associated with $-r^a$. The complex vector space structure (366), on the other hand, picks out a particular "future" time direction. Then the expectation value of $\underline{E}$ is positive provided r^a is "future-directed" in this sense. It is for this reason that we are led to select (366) as our complex vector space structure.

We summarize. If energy is to arise from time translations, there is no freedom to alter the energy operator itself. In the real Klein-Gordon case, we are forced, in order to obtain a Hilbert space, to select a preferred "future" mass shell. Then energy is positive provided r^a is future-directed. In the complex Klein-Gordon case, there are two distinct ways to obtain a Hilbert space, one which selects a preferred "future" mass shell, and one which does not. It is only the former choice which leads to positive energies. We make this choice.

There is an additional sense in which (366) is a more natural choice for the complex vector space structure for H_CKG. Every real solution of the Klein-Gordon equation is certainly also a complex solution. We thus have a mapping $\Lambda : H_{RKG} \to H_{CKG}$. This mapping is certainly norm-preserving. Is it also linear? The answer is no if we choose the structure (365), and yes if we choose the the structure (366).

A completely analogous situation holds for the other Hilbert spaces. H_M is based on real solutions of Maxwell's equations, its complex vector space structure depends on choosing a particular future mass shell, and energies are naturally positive. On the other hand, H_D and H_N are based on complex fields. We have two choices for the complex vector space structure, one of which leads to positive energies and one of which does not. We choose the complex vector space structure to be the one which, by preferring a future mass shell, makes the energy be positive. (See (320), (349).)

23 Fields as Operators: Propagators

Ordinary relativistic fields are to be replaced, eventually, by an appropriate class of operators on Fock space. This transition from fields to operators is to be carried out according to the following general rules:

i) A real field becomes a Hermitian operator; a pair of complex-conjugate fields is a pair of adjoint operators;

ii) The operators have the same index structure, and satisfy the same equations, as the corresponding fields; and

iii) The "positive-frequency part" of the operator is annihilation of a particle, the negative-frequency part creation of an antiparticle.

We have already discussed these operators in the real Klein-Gordon and Maxwell cases (Sects. 12 and 14, respectively). The purposes of this section are, firstly, to treat the complex Klein-Gordon and Dirac cases, and, secondly, to establish certain properties of the functions which appear in the commutators or anti-commutators. For completeness, we briefly review Sects. 12 and 14.

Real Klein-Gordon. H_{RKG} consists of (measurable, square-integrable) complex-valued functions $\phi(k)$ on M_μ which satisfy $\bar{\phi}(k) = \phi(-k)$. The inner product is

$$(\phi(k), \psi(k)) = \frac{1}{\hbar} \int\limits_{M_\mu^+} \phi(k)\overline{\psi(k)}\, \mathrm{d}V_\mu + \frac{1}{\hbar} \int\limits_{M_\mu^-} \overline{\phi(k)}\psi(k)\, \mathrm{d}V_\mu \tag{371}$$

Let $f(x)$ be a real test function on Minkowski space, and let $f(k)$ be its Fourier inverse, so $\bar{f}(k) = f(-k)$. Then $f(k)$, restricted to M_μ, defines an element, $\sigma(f)$, of H_{RKG}. The corresponding field operator on symmetric Fock space is

$$\underline{\phi}(f) = \hbar C(\sigma(f)) + \hbar A(\sigma(f)) \tag{372}$$

Note that (372) is Hermitian and satisfies the Klein-Gordon equation:

$$\underline{\phi}\left((\Box + \mu^2)f\right) = 0 \tag{373}$$

The commutator is

$$\begin{aligned} \left[\underline{\phi}(f), \underline{\phi}(g)\right] &= \hbar^2 \left([C(\sigma(f)), A(\sigma(g))] + [A(\sigma(f)), C(\sigma(g))]\right) \\ &= \hbar^2 \left(-\sigma^\alpha(f)\bar{\sigma}_\alpha(g) + \sigma^\alpha(g)\bar{\sigma}_\alpha(f)\right) \mathbb{I} \\ &= \frac{\hbar}{i} D(f,g)\mathbb{I} \end{aligned} \tag{374}$$

where we have defined

$$D(f,g) = -i \left(\int\limits_{M_\mu^+} - \int\limits_{M_\mu^-} \right) \left(f(k)\bar{g}(k) - \bar{f}(k)g(k)\right)\, \mathrm{d}V_\mu \tag{375}$$

Complex Klein-Gordon. H_{CKG} consists of (measurable, square-integrable) complex-valued functions $\phi(k)$ on M_μ. The inner product is

$$(\phi(k), \psi(k)) = \frac{1}{\hbar} \int\limits_{M_\mu^+} \phi(k)\overline{\psi(k)}\, \mathrm{d}V_\mu + \frac{1}{\hbar} \int\limits_{M_\mu^-} \overline{\phi(k)}\psi(k)\, \mathrm{d}V_\mu \tag{376}$$

Let $f(x)$ be a real test function on Minkowski space, and let $f(k)$ be its Fourier inverse, so $\bar{f}(k) = f(-k)$. Let $\sigma^+(f)$ be the element of H_{CKG} given by $f(k)$ on M_μ^+ and zero on M_μ^-, and let $\sigma^-(f)$ be given by $f(k)$ on M_μ^- and zero on M_μ^+. The corresponding field operators on symmetric Fock space are

$$\underline{\phi}(f) = \hbar \left(C(\sigma^-(f)) + A(\overline{\sigma^+(f)}) \right) \tag{377}$$

$$\underline{\phi}^*(f) = \hbar \left(A(\overline{\sigma^-(f)}) + C(\sigma^+(f)) \right) \tag{378}$$

Note that these are adjoints of each other, and that they satisfy the Klein-Gordon equation:

$$\underline{\phi}\left((\Box + \mu^2)f\right) = \underline{\phi}^*\left((\Box + \mu^2)f\right) = 0 \tag{379}$$

We clearly have

$$\left[\underline{\phi}(f), \underline{\phi}(g)\right] = \left[\underline{\phi}^*(f), \underline{\phi}^*(g)\right] \tag{380}$$

For the other commutator, however,

$$\begin{aligned} \left[\underline{\phi}(f), \underline{\phi}^*(g)\right] &= \hbar^2 \left(\left[C(\sigma^-(f)), A(\overline{\sigma^-(g)}) \right] + \left[A(\overline{\sigma^+(f)}), C(\sigma^+(g)) \right] \right) \\ &= \hbar^2 \left(-\sigma^{-\alpha}(f)\overline{\sigma^-_\alpha}(g) + \sigma^{+\alpha}\overline{\sigma^+_\alpha}(f) \right) \mathbb{I} = \frac{\hbar}{2i} D(f,g)\mathbb{I} \end{aligned} \tag{381}$$

where $D(f,g)$ is given by (375).

Maxwell. H_M consists of (measurable, square-integrable) complex vector functions $A_a(k)$ on M_0 which satisfy $\bar{A}_a(k) = A_a(-k)$ and $k^a A_a(k) = 0$, where two such functions which differ by a multiple of k_a are to be regarded as defining the same element of H_M. The inner product is

$$(A_a(k), B_a(k)) = -\frac{1}{\hbar} \int\limits_{M_0^+} A_a(k)\bar{B}^a(k)\, \mathrm{d}V - \frac{1}{\hbar} \int\limits_{M_0^-} \bar{A}_a(k)B^a(k)\, \mathrm{d}V \tag{382}$$

Let $f^a(x)$ be a real test vector field on Minkowski space which is the sum of a divergence-free field and a gradient, and let $f^a(k)$ be its Fourier inverse. Then $f^a(k)$ satisfies $f^a(-k) = \bar{f}^a(k)$ and $k_a f^a(k) = 0$, and so defines an element, $\sigma(f^a)$, of H_M. The corresponding operator on symmetric Fock space is

$$\underline{A}(f^a) = \hbar C(\sigma(f^a)) + \hbar A(\sigma(f^a)) \tag{383}$$

This operator is Hermitian, and satisfies Maxwell's equations (for the vector potential):

$$\underline{A}(\nabla^a f) = 0 \qquad\qquad \underline{A}(\Box f^a) = 0 \tag{384}$$

The commutator is

$$[\underline{A}(f^a), \underline{A}(g^a)] = \hbar^2 \left(-\sigma^\alpha(f^a)\bar{\sigma}_\alpha(g^a) + \sigma^\alpha(g^a)\bar{\sigma}_\alpha(f^a)\right)\mathbb{I} = \frac{\hbar}{i} D(f^a, g^a)\mathbb{I} \quad (385)$$

where we have defined

$$D(f^a, g^a) = i\left(\int\limits_{M_0^+} - \int\limits_{M_0^-}\right)\left(f_a(k)\bar{g}^a(k) - g_a(k)\bar{f}^a(k)\right)\, \mathrm{d}V \quad (386)$$

Dirac. H_D consists of (measurable, square-integrable) pairs, $(\xi^A(k), \eta_{A'}(k))$, of spinor functions on M_μ which satisfy

$$ik_{AA'}\xi^A(k) = \frac{\mu}{\sqrt{2}}\eta_{A'}(k) \qquad ik^{AA'}\eta_{A'}(k) = -\frac{\mu}{\sqrt{2}}\xi^A(k) \quad (387)$$

The inner product is

$$\left((\xi^A, \eta_{A'}), (\sigma^A, \tau_{A'})\right) = \frac{2}{\mu^2}\int\limits_{M_\mu^+} \xi^A(k)\bar{\sigma}^{A'}k_{AA'}\, \mathrm{d}V - \frac{2}{\mu^2}\int\limits_{M_\mu^-} \sigma^A(k)\bar{\xi}^{A'}k_{AA'}\, \mathrm{d}V \quad (388)$$

Let $f^A(x), \bar{f}_{A'}(x)$ be a pair of test spinor fields on Minkowski space which is real in the sense that the second field is the complex-conjugate of the first. (Of course, "test" means having compact support.) Let $f^A(k)$ be the Fourier inverse of $f^A(x)$, and consider the pair

$$\left(\mu f^A(k) - \frac{i\sqrt{2}}{\mu}k^{AA'}\bar{f}_{A'}(k), \mu\bar{f}_{A'}(k) + \frac{i\sqrt{2}}{\mu}k_{AA'}f^A(k)\right) \quad (389)$$

of spinor functions on M_μ. Note that the pair (389) satisfies (387). Let $\sigma^+(f^A, \bar{f}_{A'})$ be the element of H_D which is given by (389) on M_μ^+ and zero on M_μ^-, and let $\sigma^-(f^A, \bar{f}_{A'})$ be the element given by (389) on M_μ^- and zero on M_μ^+. The corresponding operators on antisymmetric Fock space are

$$\underline{\psi}(f^A, \bar{f}_{A'}) = C(\sigma^-(f^A, \bar{f}_{A'})) + A(\overline{\sigma^+(f^A, \bar{f}_{A'})}) \quad (390)$$

$$\underline{\psi}^*(f^A, \bar{f}_{A'}) = A(\overline{\sigma^-(f^A, \bar{f}_{A'})}) + C(\sigma^+(f^A, \bar{f}_{A'})) \quad (391)$$

These operators are adjoints of each other, and satisfy the Dirac equation:

$$\underline{\psi}\left(\mu f^A(k) - \frac{\sqrt{2}}{\mu}\nabla^{AA'}\bar{f}_{A'}(k), \mu\bar{f}_{A'}(k) + \frac{\sqrt{2}}{\mu}\nabla_{AA'}f^A(k)\right)$$
$$= \underline{\psi}^*\left(\mu f^A(k) - \frac{\sqrt{2}}{\mu}\nabla^{AA'}\bar{f}_{A'}(k), \mu\bar{f}_{A'}(k) + \frac{\sqrt{2}}{\mu}\nabla_{AA'}f^A(k)\right) = 0 \quad (392)$$

We clearly have

$$\left\{\underline{\psi}(f^A, \bar{f}_{A'}), \underline{\psi}(g^A, \bar{g}_{A'})\right\} = \left\{\underline{\psi}^*(f^A, \bar{f}_{A'}), \underline{\psi}^*(g^A, \bar{g}_{A'})\right\} = 0 \tag{393}$$

For the other anticommutator,

$$\begin{aligned}
&\left\{\underline{\psi}(f^A, \bar{f}_{A'}), \underline{\psi}^*(g^A, \bar{g}_{A'})\right\} \\
&\quad = \left(\sigma^{-\alpha}(f^A, \bar{f}_{A'})\bar{\sigma}^-_\alpha(g^A, \bar{g}_{A'}) + \sigma^{+\alpha}(g^A, \bar{g}_{A'})\bar{\sigma}^+_\alpha(f^A, \bar{f}_{A'})\right)\mathbb{I} \\
&\quad = D\left((f^A, \bar{f}_{A'}), (g^A, \bar{g}_{A'})\right)\mathbb{I}
\end{aligned} \tag{394}$$

where, using (388) and (389),

$$\begin{aligned}
&D\left((f^A, \bar{f}_{A'}), (g^A, \bar{g}_{A'})\right) \\
&\qquad = 2\left(\int\limits_{M^+_\mu} - \int\limits_{M^-_\mu}\right)\Big[\Big(f^A(k)\bar{g}^{A'}(k) + \bar{f}^{A'}(k)g^A(k)\Big)\, k_{AA'} \\
&\qquad\qquad + \frac{i\mu}{\sqrt{2}}\Big(f^A(k)g_A(-k) - \bar{f}^{A'}(k)\bar{g}_{A'}(-k)\Big)\Big]\, \mathrm{d}V
\end{aligned} \tag{395}$$

The functions (375), (386), and (395) play a very special role in relativistic quantum field theory. They are called *Feynman propagators*. Several properties of the propagators follow immediately from the definitions. In the first place, they are all real. Secondly, we have the symmetries

$$D(f, g) = -D(g, f) \tag{396}$$

$$D(f^a, g^a) = -D(g^a, f^a) \tag{397}$$

$$D\left((f^A, \bar{f}_{A'}), (g^A, \bar{g}_{A'})\right) = D\left((g^A, \bar{g}_{A'}), (f^A, \bar{f}_{A'})\right) \tag{398}$$

Furthermore, since the propagators arise from commutators or anticommutators of the field operators, they satisfy the appropriate field equations:

$$D\left((\Box + \mu^2)f, g\right) = 0 \tag{399}$$

$$D(\nabla^a f, g^a) = 0 \qquad D(\Box f^a, g^a) = 0 \tag{400}$$

$$D\left(\left(\mu f^A(k) - \frac{\sqrt{2}}{\mu}\nabla^{AA'}\bar{f}_{A'}(k), \mu\bar{f}_{A'}(k) + \frac{\sqrt{2}}{\mu}\nabla_{AA'}f^A(k)\right), (g^A, \bar{g}_{A'})\right) = 0 \tag{401}$$

Note also that the propagators are linear in the real test fields on which they depend.

A more remarkable property of the propagators is that they can all be expressed directly in terms of the Klein-Gordon propagator, $D(f, g)$. Let v^a and w^a be constant vector fields on Minkowski space, f and g real test functions, and consider the expression

$$-v^a w_a D(f, g) \tag{402}$$

Inserting (375) in (402), we see that (402) is precisely the right side of (386), provided we set $f^a = fv^a$, $g^a = wg^a$. Thus, we may define a function $D(f^a, g^a)$ for test fields of the form $f^a = fv^a$, $g^a = wg^a$ by (402). Then, assuming linearity in f^a and g^a, we extend the range of $D(f^a, g^a)$ to arbitrary real test vector fields. Finally, restricting f^a and g^a to fields which can be expressed as the sum of a divergence-free field and a gradient, we obtain precisely the Maxwell propagator.

The situation is similar in the Dirac case. Let ξ^A and σ^A be constant spinor fields, f and g real test functions. Then, from (395) and (375), we have

$$D\left((f^A, \bar{f}_{A'}), (g^A, \bar{g}_{A'})\right) = 2D\left((\xi^A\bar{\sigma}^{A'} + \sigma^A\bar{\xi}^{A'})\nabla_{AA'}f, g\right) + \frac{\mu}{\sqrt{2}}(\xi^A\sigma_A + \bar{\xi}^{A'}\bar{\sigma}_{A'})D(f, g) \quad (403)$$

But, again by linearity, if we know the Dirac propagator for test fields of the form $f(\xi^A, \bar{\xi}_{A'})$ with constant ξ^A, we know the Dirac propagator for all test fields. In this sense, then, the Dirac propagator follows already from the Klein-Gordon propagator.

We may now derive a particularly important property of the Feynman propagators. A function of a pair of test fields will be called *causal* if it vanishes whenever the test fields have relatively spacelike supports (see p. 39). We have seen in Sect. 12 that the Klein-Gordon propagator, $D(f, g)$, is causal. The remarks above imply, therefore, that all the Feynman propagators are causal.

24 Spin and Statistics

What is it that determines whether the appropriate Fock space for an elementary particle is the symmetric or the antisymmetric one? (This distinction is said to one of *statistics*. Particles described by the symmetric Fock space are called *bosons*, and are said to satisfy *Bose statistics*. Particles described by the antisymmetric Fock space are called *fermions*, and are said to satisfy *Fermi statistics*.) It is found in Nature that the statistics a particle obeys is invariably correlated with another feature of the particle, its spin. It is found, in fact, that all particles with half-integer (i.e., half-odd-integer) spin obey Fermi statistics, while particles with integer spin obey Bose statistics. How should this fact be incorporated into quantum field theory? One could, of course, merely regard the correlation between spin and statistics as an empirical fact — a fact which can be used to choose the appropriate statistics in each case. It is natural to ask, however, whether there is some deeper theoretical reason why Nature operates as She does. Certainly, no obvious internal inconsistencies arise if we insist that Klein-Gordon and Maxwell particles be fermions, while Dirac particles be bosons. It would be desirable, however, to find some very general requirement on quantum field theories which would force the experimentally observed relation between spin and statistics. There is, in fact, such a requirement: the demand that the propagators be causal. We have seen in Sect. 23 that, with the

"correct" statistics, the propagators are indeed causal. In this section, we shall indicate why the propagators for "fermion Klein-Gordon", "fermion Maxwell", and "boson Dirac" particles are not causal. These results are a special case of a core general theorem. If we require that energy be positive (to fix the complex vector space structure), and that the propagator be causal, then particles with half-integer spin must be fermions and those with integer spin bosons. We shall not discuss this general theorem further.

We begin with H_{RKG}. The (one-particle) Hilbert space is the same as before, the inner product given by (371). The operator $\underline{\phi}(f)$ is still defined by (372). Now, however, we suppose that the creation and annihilation operators act on antisymmetric Fock space. Then (373) still holds, but (374) must be modified as follows:

$$\begin{aligned}\left\{\underline{\phi}(f), \underline{\phi}(g)\right\} &= \hbar^2\{C(\sigma(f)), A(\sigma(g)\} + \hbar^2\{A(\sigma(f)), C(\sigma(g))\} \\ &= \hbar^2 \left(\sigma^\alpha(f)\bar{\sigma}_\alpha(g) + \sigma^\alpha\bar{\sigma}_\alpha(f)\right)\mathbb{I} \\ &= \hbar \left(\int\limits_{M_\mu^+} + \int\limits_{M_\mu^-}\right)\left(f(k)\bar{g}(k) + \bar{f}(k)g(k)\right)\, \mathrm{d}V\, \mathbb{I}\end{aligned} \tag{404}$$

The "propagator" for antisymmetric statistics — the last line in (404) — is simply not causal. (Proof: Choose almost any test functions f and g with relatively spacelike supports, and evaluate the integral.) That is to say, we obtain a causal propagator in the real Klein-Gordon case if and only if we use Bose statistics. Thus, if we take causality of the propagator as a fundamental assumption, we are led to assign Bose statistics to real Klein-Gordon particles.

Now consider the complex Klein-Gordon case. If we choose Fermi statistics, (376), (377), (378), and (379)) still hold. Furthermore, (380) holds if we replace the commutators by anticommutators. For (381), however, we have

$$\begin{aligned}\left\{\underline{\phi}(f), \underline{\phi}^*(g)\right\} &= \hbar^2\{C(\sigma^-(f)), A(\overline{\sigma^-(g)})\} + \hbar^2\{A(\overline{\sigma^+(f)}), C(\sigma^+(g))\} \\ &= \hbar^2 \left(\sigma^{-\alpha}(f)\bar{\sigma}^-_\alpha(g) + \sigma^{+\alpha}(g)\bar{\sigma}^+_\alpha(f)\right)\mathbb{I} \\ &= \frac{\hbar}{2}\left(\int\limits_{M_\mu^+} + \int\limits_{M_\mu^-}\right)\left(f(k)\bar{g}(k) + \bar{f}(k)g(k)\right)\, \mathrm{d}V\, \mathbb{I}\end{aligned} \tag{405}$$

But the last line of (405) is not causal. Hence, in order to obtain a causal propagator, complex Klein-Gordon particles must be bosons.

If we assign Fermi statistics to H_M, (382), (383), and (384) remain valid. But (385) becomes

$$\begin{aligned}\{\underline{A}(f^a), \underline{A}(g^a)\} &= \hbar^2 \left(\sigma^\alpha(f^a)\bar{\sigma}_\alpha(g^a) + \sigma^\alpha(g^a)\bar{\sigma}_\alpha(f^a)\right)\mathbb{I} \\ &= -\hbar \left(\int\limits_{M_\mu^+} + \int\limits_{M_\mu^-}\right)\left(f^a(k)\bar{g}_a(k) + \bar{f}^a(k)g^a(k)\right)\, \mathrm{d}V\, \mathbb{I}\end{aligned} \tag{406}$$

The last line is not causal. So causality of the propagator implies Bose statistics for photons.

Finally, we attempt to impose Bose statistics on Dirac particles. Eqns. (387), (388), (389), (390), (391), and (392) remain valid. Eqn. (393) remains valid if the anticommutators are replaced by commutators. But (394) becomes

$$\begin{aligned}
&\left[\underline{\psi}(f^A, \bar{f}_{A'}), \underline{\psi}^*(g^A, \underline{g}_{A'})\right] \\
&\quad = \left(-\sigma^{-\alpha}(f^A, \bar{f}_{A'})\bar{\sigma}^-_\alpha(g^A, \bar{g}_{A'}) + \sigma^{+\alpha}(g^A, \bar{g}_{A'})\bar{\sigma}^+_\alpha(f^A, \bar{f}_{A'})\right)\mathbb{I} \\
&\quad = -2\left(\int\limits_{M^+_\mu} + \int\limits_{M^-_\mu}\right)\left[\left(f^A(k)\bar{g}^{A'}(k) - \bar{f}^{A'}(k)g^A(k)\right)k_{AA'}\right. \\
&\qquad\qquad \left. + \frac{i\mu}{\sqrt{2}}\left(\bar{f}^{A'}(k)\bar{g}_{A'}(-k) + f^A(k)g_A(-k)\right)\right]\,\mathrm{d}V\,\mathbb{I}
\end{aligned} \tag{407}$$

which, again, is not causal. Causality of the propagator implies Fermi statistics for Dirac particles.

We summarize. The requirement that energies be positive fixes the complex vector space structure of the one-particle Hilbert spaces. The additional requirement that the propagators (the commutators or anticommutators of the field operators) be causal then requires that particles with integer spin be bosons and particles with half-integer spin be fermions, at least for the our cases H_{RKG}, H_{CKG}, H_M, and H_D.

25 $\star$-Algebras

We have now obtained a number of quantum field theories of relativistic, noninteracting particles. Our approach consists, basically, of the following steps:

i) form a Hilbert space of an appropriate collection of solutions of the field equations (Klein-Gordon, Maxwell, Dirac, neutrino),

ii) introduce the corresponding symmetric or antisymmetric Fock space, and

iii) replace the original fields by operators on Fock space.

However, there exists an alternative approach, in which one begins with the field operators and their commutators (or anticommutators) as the basic objects, deriving from these the Fock space and finally the one-particle Hilbert space. While the two approaches are completely equivalent logically, they differ considerably in attitude. In particular, the alternative approach emphasizes the analogy between second quantization (the ultimate passage from fields to field operators) and first quantization (e.g., the passage from Newtonian mechanics to the Schrödinger equation). One thinks of the fields (Klein-Gordon, Maxwell, Dirac, etc.) as "classical quantities" (analogous to x and p in Newtonian

mechanics) which, in the quantized version of the theory, are to become operators on some Hilbert space. This alternative approach is the one conventionally followed in textbooks. We discuss it in this section.

It is convenient to first introduce a mathematical object. A $\star$-*algebra* consists, first of all, of an associative algebra $\mathcal{A}$ (over the complexes) with unit $\mathbb{I}$. That is to say, $\mathcal{A}$ is a complex vector space on which there is defined a product, AB, between elements A and B of $\mathcal{A}$, where this product is linear in the factors and satisfies (254). Furthermore, there is an element $\mathbb{I}$ of $\mathcal{A}$ such that

$$\mathbb{I}A = A\mathbb{I} = A \tag{408}$$

for every $A \in \mathcal{A}$. (Clearly, this $\mathbb{I}$ is unique.) Furthermore, we require that, as part of the structure of a $\star$-algebra, there be given a mapping from $\mathcal{A}$ to $\mathcal{A}$ (the image of $A \in \mathcal{A}$ under this mapping written $A^\star$) subject to:

A1. For each $A \in \mathcal{A}$, $(A^\star)^\star = A$.

A2. For each $A, B \in \mathcal{A}$, $(AB)^\star = B^\star A^\star$.

A3. For each $A, B \in \mathcal{A}$, $\alpha \in \mathbb{C}$, $(\alpha A + B)^\star = \bar{\alpha} A^\star + B^\star$.

The standard example of a $\star$-algebra is the collection of all bounded operators which are defined everywhere on a fixed Hilbert space H. Addition is defined by adding the operators, and scalar multiplication by multiplying the operator by the complex number in the usual way. The product of two operators is defined by applying them in succession. The unit $\mathbb{I}$ is, of course, the identity operator. Finally, "$\star$" represents the operation of taking the adjoint of an operator. (In fact, every bounded operator defined everywhere on a Hilbert space H has a unique adjoint. We omit the (moderately difficult) proof.) Note that it is well-defined to speak of projection, Hermitian, and unitary elements of a $\star$-algebra, for these notions involve only the structure incorporated in to the $\star$-algebra. Intuitively, one can think of a $\star$-algebra as representing "operators on a Hilbert space, but without the Hilbert space itself."

The essential idea of the approach to be described below is identical for all the relativistic field equations. It will suffice, therefore, to treat one case in detail. We select the complex Klein-Gordon fields.

The idea is to first introduce a certain $\star$-algebra $\mathcal{A}$. We suppose that, with each real test function f on Minkowski space, there is associated a pair of elements of $\mathcal{A}$, $\underline{\phi}(f)$ and $\underline{\phi} \star (f)$, which are related by the $\star$-operation. We suppose, furthermore, that $\mathcal{A}$ is generated by $\mathbb{I}$, $\underline{\phi}(f)$, and $\underline{\phi} \star (f)$ (as f runs over all test functions). That is to say, the most general element of $\mathcal{A}$ consists of a finite linear combination, with complex coefficients, of $\mathbb{I}$ and products of the $\underline{\phi}(f)$'s and $\underline{\phi} \star (f)$'s, e.g.,

$$\alpha\mathbb{I} + \beta\underline{\phi}^\star(f) + \gamma\underline{\phi}(g)\underline{\phi}^\star(k) + \delta\underline{\phi}(m)\underline{\phi}^\star(n)\underline{\phi}(p) \tag{409}$$

Clearly, we can take the sum or product of objects of the form (409) multiply such an object by a complex number, and take the $\star$ of such an object. Unfortunately, we still do not have quite the $\star$-algebra we require. We wish to

require, in addition, that certain expressions of the form (409), while formally distinct, are to be regarded as equal as elements of $\mathcal{A}$. That is to say, we wish to impose certain relations among the elements (409) of $\mathcal{A}$. (This construction is analogous to that in which one obtains a group by postulating the existence of certain elements subject to relations. If we wished to be more formal, we would introduce an equivalence relation.) We impose the following relations:

$$\underline{\phi}(af+g) = a\underline{\phi}(f) + \underline{\phi}(g) \qquad \underline{\phi}^\star(af+g) = a\underline{\phi}^\star + \underline{\phi}^\star(g) \tag{410}$$

$$\underline{\phi}\left((\Box+\mu^2)f\right) = \underline{\phi}^\star\left((\Box+\mu^2)f\right) = 0 \tag{411}$$

$$\left[\underline{\phi}(f), \underline{\phi}(g)\right] = \left[\underline{\phi}^\star(f), \underline{\phi}^\star(g)\right] = 0 \tag{412}$$

$$\left[\underline{\phi}(f), \underline{\phi}^\star(g)\right] = \frac{\hbar}{2i} D(f,g)\,\mathbb{I} \tag{413}$$

where f and g are any test functions, a is any real number, and $D(f,g)$ is the Feynman propagator, (375). This completes the specification of the $\star$-algebra $\mathcal{A}$. (Although, of course, this $\star$-algebra $\mathcal{A}$ looks familiar, it is to be regarded, for the present as merely the mathematical object which results from the construction above.)

It is useful conceptually to restate the construction above from a more physical point of view. We have taken the classical field $\phi(x)$ and its complex-conjugate field $\bar{\phi}(x)$, and replaced them by operators, $\underline{\phi}(f)$ and $\underline{\phi}^\star(f)$. ("Operators" which, as yet, act on no particular Hilbert space: therefore, elements of a $\star$-algebra.) We impose on these operators a number of more or less natural conditions. We require that the operators be linear in the test functions, (410). We require that the operators, in their position dependence, satisfy the same equations as the fields they replaced, (411). We require that the $\underline{\phi}$'s commute with each other, and that the $\underline{\phi}^\star$'s commute with each other, (412). (This condition is analogous to the statement in Schrödinger theory that position operators commute with each other and momentum operators commute with each other.) We must be careful, however, that we do not have all the operators commute with each other, for the passage from classical to quantum theory involves replacing classical variables by operators with certain, nontrivial, commutation relations. What should we choose for the commutator of $\underline{\phi}(f)$ and $\underline{\phi}(g)$? We require, firstly, that the commutator be a multiple of the identity (just as the commutator of the operators x and p in the Schrödinger theory is a multiple of the identity). We require, furthermore, that $[\underline{\phi}(f), \underline{\phi}^\star(g)]$ vanish when f and g have relatively spacelike supports. (This assumption is perhaps not too unreasonable physically. When f and g have relatively spacelike supports, then, by causality, measurements made in the support of $\mathbb{I}$ should in no way affect measurements made in the support of g. Since the field operators are, in some sense, to be associated with measurements, we might expect commutativity in this case.) But these conditions ((410), (411), (412), and the assumption that the right side of (413) be a multiple of the identity which is a causal function) imply that the right side of (413) be precisely the Feynman propagator, up to an

overall factor. (This statement is not difficult to prove.) Note that the "passage to a quantum theory" arises because of the assumed noncommutativity, (413).

To summarize, if we take the Klein-Gordon equation as a "classical" equation, and attempt to "quantize" it, more or less in the standard way, we are led to introduce the $\star$-algebra $\mathcal{A}$.

We next construct an inner-product space. While the construction could in principle be carried out directly in terms of the $\star$-algebra $\mathcal{A}$, it is more convenient to first introduce a second $\star$-algebra $\mathcal{B}$. This $\mathcal{B}$ is to be generated by elements $A^+(f)$, $C^+(f)$, $A^-(f)$, and $C^-(f)$, for each real test function f, subject to the following relations:

$$(A^+(f))^\star = C^+(f) \qquad (A^-(f))^\star = C^-(f) \tag{414}$$

$$\begin{aligned} A^\pm(af+g) &= aA^\pm(f) + A^\pm(g) \\ C^\pm(af+g) &= aC^\pm(f) + C^\pm(g) \end{aligned} \tag{415}$$

$$A^\pm\left((\Box+\mu^2)f\right) = C^\pm\left((\Box+\mu^2)f\right) = 0 \tag{416}$$

$$\begin{gathered} \big[C^+(f), C^+(g)\big] = \big[C^+(f), C^-(g)\big] = \big[C^-(f), C^-(g)\big] = 0 \\ \big[A^+(f), A^+(g)\big] = \big[A^+(f), A^-(g)\big] = \big[A^-(f), A^-(g)\big] = 0 \\ \big[A^+(f), C^-(g)\big] = \big[A^-(f), C^+(g)\big] = 0 \end{gathered} \tag{417}$$

$$\big[A^+(f), C^+(g)\big] = \frac{\hbar}{2i} D^+(f,g)\mathbb{I} \qquad \big[A^-(f), C^-(g)\big] = \frac{\hbar}{2i} D^-(f,g)\mathbb{I} \tag{418}$$

First note that we may regard $\mathcal{A}$ as a $\star$-subalgebra of $\mathcal{B}$. Specifically, we set

$$\begin{aligned} \underline{\phi}(f) &= \hbar\big[C^-(f) + A^+(f)\big] \\ \underline{\phi}^\star(f) &= \hbar\big[C^+(f) + A^-(f)\big] \end{aligned} \tag{419}$$

whence each element of $\mathcal{A}$ defines an element of $\mathcal{B}$. Note that the identifications (419) indeed establish $\mathcal{A}$ as $\star$-subalgebra of $\mathcal{B}$, for (414)–(418) imply (410)–(413). In fact, although $\mathcal{B}$ is larger than $\mathcal{A}$, there is a sense in which $\mathcal{B}$ "does not add anything new" to $\mathcal{A}$. Specifically, each element of $\mathcal{B}$ can be considered as limiting case of elements of $\mathcal{A}$: $A^+(f)$, for example, is the "positive-frequency part" of $\underline{\phi}(f)$. (Just as a complex-valued solution $\phi(x)$ of the Klein-Gordon equation can be decomposed into its positive-frequency and negative-frequency parts, so can an operator-valued solution, $\phi(f)$, of the Klein-Gordon equation. It in perhaps not surprising that if one introduces enough machinery, it becomes possible to describe such a decomposition directly in terms of the $\star$-algebra.)

Why do we introduce two distinct $\star$-algebras when they carry essentially the same information? Because $\mathcal{A}$ is easier to motivate while $\mathcal{B}$ is easier to manipulate.

We now construct our inner-product space, $K_\mathcal{B}$. We postulate, first of all, the existence of an element σ_0 (the "vacuum", more commonly written $|0\rangle$). The most general element of $K_\mathcal{B}$ is to consist of the juxtaposition of an element

of $\mathcal{B}$ and σ_0. We add such elements of $K_{\mathcal{B}}$, and multiply by complex numbers, by performing the indicated operation on $\mathcal{B}$, e.g.,

$$\alpha(A\sigma_0) + (B\sigma_0) = (\alpha A + B)\sigma_0 \tag{420}$$

where $A, B \in \mathcal{B}$, $\alpha \in \mathbb{C}$. We wish, however, to impose on these elements a further relation, namely

$$A^+(f)\sigma_0 = 0 \qquad A^-(f)\sigma_0 = 0 \tag{421}$$

("annihilation on the vacuum gives zero"). We now have a complex vector space. To obtain an inner-product space, we must introduce a norm. To evaluate the norm of an element of $K_{\mathcal{B}}$, one first formally takes the "inner product" of the element with itself. For $\alpha A^+(f)C^-(g)C^+(h)\sigma_0$, for example, one would write

$$(\alpha A^+(f)C^-(g)C^+(h)\sigma_0, \alpha A^+(f)C^-(g)C^+(h)\sigma_0) \tag{422}$$

We now set down certain rules for manipulating such expressions. Firstly, an element of $\mathcal{B}$ which appears first on either side of the "formal inner product" can be transferred to the other side (where it must also appear first) provided that, simultaneously, it is replaced by its starred version. (That is, we mimic the usual rule for transferring an operator to the other side of an inner product.) For example, (423) can be rewritten

$$(\alpha C^-(g)C^+(h)\sigma_0, \alpha C^+(f)A^+(f)C^-(g)C^+(h)\sigma_0) \tag{423}$$

or

$$(\alpha\bar{\alpha}A^-(g)C^+(f)A^+(f)C^-(g)C^+(h)\sigma_0, C^+(h)\sigma_0) \tag{424}$$

Secondly, one can use the commutation relations (417) and (418). For example, (423) can be rewritten

$$\begin{aligned}(\alpha C^-(g)C^+(h)\sigma_0, \alpha A^+(f)C^+(f)C^-(g)C^+(h)\sigma_0) \\ + (\alpha C^-(g)C^+(h)\sigma_0, \alpha\frac{\hbar}{2i}D^+(f,f)C^-(g)C^+(h)\sigma_0)\end{aligned} \tag{425}$$

Thirdly, one can use (421). Finally, we postulate

$$(\sigma_0, \sigma_0) = 1 \tag{426}$$

("the vacuum is normalized to unity"). By using these rules, every "norm" can be reduced to some number. This is done, roughly speaking, as follows. First use the commutators to "push the annihilation operators to the right" until they stand next to σ_0 and hence give zero. There then remain only creation operators. Each of these, in turn, is transferred to the other side of the inner product, thus becoming an annihilation operator. Each annihilation operator obtained in this way is then "pushed to the right" again, where it eventually meets σ_0 and gives zero. In this way, all the operators are eventually eliminated,

leaving only the functions which appear in the commutators. Now use (426). As a simple example, we evaluate the norm of $C^+(f)\sigma_0$:

$$\begin{aligned}(C^+(f)\sigma_0, C^+(f)\sigma_0) &= (A^+(f)C^+(f)\sigma_0, \sigma_0) \\ &= (C^+(f)A^+(f)\sigma_0, \sigma_0) + (\frac{\hbar}{2i}D^+(f,f)\sigma_0, \sigma_0) \\ &= 0 + \frac{\hbar}{2i}D^+(f,f)(\sigma_0, \sigma_0) \\ &= \frac{\hbar}{2i}D^+(f,f)\end{aligned} \tag{427}$$

Thus, $K_{\mathcal{B}}$ has the structure of an inner-product space. (We shall establish shortly that the norm, defined above, is indeed positive.) In particular, if we consider only the elements of $K_{\mathcal{B}}$ which can be obtained by applying elements of $\mathcal{A}$ to σ_0 we obtain an inner-product subspace, $K_{\mathcal{A}}$, of $K_{\mathcal{B}}$. Finally, we take the completion, $\overline{K_{\mathcal{A}}}$, of $K_{\mathcal{A}}$ to obtain a Hilbert space. (In fact, $K_{\mathcal{A}}$ is dense in $K_{\mathcal{B}}$, so $\overline{K_{\mathcal{A}}} = \overline{K_{\mathcal{B}}}$.)

All these formal rules and relations sound rather mysterious. It is easy, however, to see what the resulting Hilbert space is. Consider the symmetric Fock space based on H_{CKG}. As we have mentioned, the $\star$-algebras $\mathcal{A}$ and $\mathcal{B}$ can be represented as operators on this Hilbert space. Consider now the element $(1, 0, 0, \ldots)$ (see (98)) of Fock space. It satisfies (421) and (426). Clearly, the inner-product space $K_{\mathcal{A}}$ (resp. $K_{\mathcal{B}}$) is identical to the inner-product space consisting of all elements of Fock space which can be obtained by applying elements of $\mathcal{A}$ (resp. $\mathcal{B}$) to $(1, 0, 0, \ldots)$. Thus, $K_{\mathcal{A}}$ and $K_{\mathcal{B}}$ can be considered as subspaces of our Fock space. But in fact, both these subspaces are dense in Fock space. Hence, $\overline{K_{\mathcal{A}}}$ and $\overline{K_{\mathcal{B}}}$ are identical with symmetric Fock space based on H_{CKG}. In other words, we have simply re-obtained Fock space by a different route.

We summarize the situation. In the conventional approach, one begins with the classical fields, which are replaced by "operators" (elements of a $\star$-algebra), subject to certain commutation relations. One then assumes the existence of a vacuum, and builds an inner-product space by applying the elements of our $\star$-algebra to the vacuum. The norm is defined by formal manipulative rules, using the postulated commutators. Finally, one completes this inner-product space to obtain a Hilbert space. One has the feeling that one is "quantizing" a classical theory. We have proceeded in a rather different direction. We "looked ahead" to see what the resulting Hilbert space would be, and simply wrote it out. It turned out to be what we called the symmetric Fock based on the Hilbert space H_{CKG}, which, in turn, was based on the solutions of the original equation. We then simply defined the action of creation operators, annihilation operators, and field operators on this explicit Fock space. The resulting mathematical structures are identical — the methods of deriving this structure quite different. We have sacrificed much of the motivation to gain a certain explicitness.

26 Scattering: The S-Matrix

Our discussion so far has been restricted to free particles. That is to say, we have been dealing with systems consisting of any number of identical particles which interact neither with themselves nor with particles of any other type. While such systems provide a convenient starting point for quantum field theory, they are by themselves of very little physical interest. Particles in the real world do interact with other particles: electrons have charge, and so interact with photons; nucleons interact, at least, with π-mesons to form nuclei, etc. Furthermore, even for systems in which interactions play a minor role, the experiments which provide information about such systems must come from the interaction with other systems (i.e., the experimental apparatus). One of the most important situations — from both the theoretical and experimental points of view — in which interactions play a significant role is that of scattering. In this section, we shall set up the general framework for the description of scattering experiments.

We first recall a general principle of quantum theory. Let S_1 and S_2 represent two quantum systems which, we assume, in no way interact with each other. Let H_1 be the Hilbert space which encompasses the possible states of S_1, and H_2 the Hilbert space for S_2. It is because the systems do not interact that each is characterized by its own Hilbert space. Now suppose we introduce a new quantum system, S, which consists of S_1 and S_2 together. Note that we are not here turning on any interactions — we have merely decided to consider two systems as a single system. What is the Hilbert space of states of the system S? It is $H_1 \otimes H_2$. (Note: the tensor product, not the direct sum.) That is to say, a state of S can be obtained by taking a formal sum of formal products of states of S_1 and S_2. (Simple example: if H_1 and H_2 were both one-dimensional, so S_1 and S_2 each had essentially one state, then S should also have essentially one state. But in this example $H_1 \oplus H_2$ is two-dimensional, whereas $H_1 \otimes H_2$ is one-dimensional.) Note, incidentally, that any operator on H_1 (i.e., which acts on S_1) extends naturally to an operator on $H_1 \otimes H_2$ (i.e., extends to an operator which acts on S).

Now suppose we wish to consider a situation in which only certain types of particles will be permitted to interact — say, electrons-positrons, photons, and neutral π-mesons. We begin by writing down the Hilbert space $\mathcal{H}$ which encompasses the states of such a system when the interactions are "turned off". That is to say, we imagine a system in which our various particles co-exist but do not interact, and describe its states by $\mathcal{H}$. In our example, $\mathcal{H}$ would be the tensor product of the antisymmetric Fock space based on H_D, the symmetric Fock space based on H_M, and the symmetric Fock space based on H_{RKG}. Note that this is a purely mathematical construct. In the real world such particles would interact: we do not have the option of turning off interactions to suit our convenience.

It is in terms of this $\mathcal{H}$ that scattering processes are described. We consider the following situation. In the distant past, our particles are represented by broad and widely separated wave packets. These particles then enter a region in which the amplitudes are large and the wave packets overlap significantly.

They interact. Finally, we suppose that, in the distant future, all the particles are again widely separated. We wish to describe such a process as follows. It is perhaps not unreasonable to suppose that, as one goes further and further into the past, the interactions play a smaller and smaller role. Thus, in the limit as one goes into the past, it might be possible to describe the system by an element of our non-interacting Hilbert space $\mathcal{H}$. That is, our incoming state is to be some element of $\mathcal{H}$. Similarly, in the limit as one goes into the future, i.e., for the outgoing state, one obtains some other element of $\mathcal{H}$. It is only in these "past and future limits" that $\mathcal{H}$ provides an accurate description of the state of the system. In any actual finite region of Minkowski space — and particularly where the interactions are strong — a description in terms of $\mathcal{H}$ is impossible. We therefore simply abandon, for the time being, any attempt to describe in detail what is happening while the interaction takes place. We agree that all we care to know about the interaction is simply the relation between the incoming state and the outgoing state — both elements of $\mathcal{H}$. This relation is given by some mapping S from $\mathcal{H}$ to $\mathcal{H}$.

We illustrate this idea with a classical analogy. (Caution: This analogy can be misleading if pushed too far.) Suppose we are interested in solutions of

$$(\Box + \mu^2)\phi = \phi^3 \tag{428}$$

Let $\mathcal{L}$ denote the collection of solutions of this equation which are, in some suitable sense, well-behaved asymptotically. (Note that $\mathcal{L}$ is not a vector space. It is analogous to the states of the interacting quantum system, which do form a vector space.) Let $\mathcal{H}$ denote the collection of asymptotically well-behaved solutions of the Klein-Gordon equation. Fix a time-coordinate t in Minkowski space (i.e., $\nabla_a t$ is constant, unit, and timelike). For each value of t, we define a mapping, $\Lambda(t)$, from $\mathcal{L}$ to $\mathcal{H}$. Fix t_0. Then, if $\phi(x)$ is a solution of (428), the values of ϕ and $(\nabla^a t)\nabla_a phi$ on the 3-surface $t = t_0$ are initial data for some solution of the Klein-Gordon equation, which we write $\Lambda(t)\phi$. Clearly, this mapping $\Lambda(t)$ is invertible. We now ask whether the right side of

$$S = \lim_{\substack{t_2 \to \infty \\ t_1 \to -\infty}} \Lambda(t_2)\Lambda(t_1)^{-1} \tag{429}$$

exists and is independent of our original choice of time-coordinate. If so, we obtain a mapping S from $\mathcal{H}$ to $\mathcal{H}$. This S clearly provides a great deal of information about the structure of Eqn. (428).

We now return to quantum field theory. All the information we want about the interactions is to be contained in the mapping S, called the *S-matrix*, from the non-interacting Hilbert space to itself. One could, of course, merely determine S, as best as one can, from experiments. But this would hardly represent a physical theory. Ultimately, we shall be concerned with the problem of calculating S from specific assumptions concerning the nature of the interaction. It is of interest, however, to first ask whether there are any very general properties of S which one might expect to hold merely from its physical interpretation. In fact, there are two such properties. The first is that S is an operator on $\mathcal{H}$, i.e., S

is linear. I do not know of a water-tight physical argument for this assumption. It is, however, suggested by the principle of superposability in quantum theory. Let σ_1 and σ_2 be unit, orthogonal elements of H. Then $\sigma = (\sigma_1 + \sigma_2)/\sqrt{2}$ is also a unit vector in $\mathcal{H}$. A system whose incoming state is σ has probability $1/2$ that its incoming state is σ_1, and probability $1/2$ that its incoming state is σ_2. Hence, we might expect the corresponding outgoing state, $S(\sigma)$, to be the same linear combination of $S(\sigma_1)$ and $S(\sigma_2)$, i.e., we might expect to have

$$S(\sigma) = \frac{1}{\sqrt{2}} \left(S(\sigma_1) + S(\sigma_2) \right) \tag{430}$$

These considerations strongly suggest the assumption we now make: that S is a linear operator on $\mathcal{H}$. The second property of S follows from the probabilistic interpretation of states in quantum theory. Let σ be a unit vector in $\mathcal{H}$. Then, if we write $\mathcal{H}$ as a direct sum of certain of its orthogonal subspaces, the sum of the norms of the projections of σ into these subspaces is one. This fact is interpreted as meaning that the total probability of the system's being found in one of these subspaces is one. But if this σ is our incoming state, the total probability for all possible outgoing states must also be one. Hence, we might expect to have

$$\|S\sigma\| = 1 \tag{431}$$

provided $\|\sigma\| = 1$. In other words, we expect S to be a unitary operator on $\mathcal{H}$.

To summarize, the probabilities for all possible outcomes of all possible scattering experiments (involving a certain, given list of particles) are specified completely by a unitary operator S on the non-interacting Hilbert space $\mathcal{H}$. We want to find this S.

The S-matrix approach to scattering problems involves a number of physical assumptions. Among these are the following:

1. In the limit to the distant past (and distant future), the interactions have negligible influence, so the state of the actual physical system can be associated, in these limits, with elements of the non-interacting Hilbert space, $\mathcal{H}$.

2. The interaction is completely described by the S-matrix (e.g., there are no bound states.)

3. One can find short list of particles such that, if only particles which appear on this list are involved in the incoming state, then all outgoing particles will also appear on the list.

In fact, all of these assumptions are believed to be false:

1. Even in the distant past and future, particles carry a "cloud of virtual particles" which affect, for example, the observed mass. Thus, the interactions are important even in the limits. It appears, however, that these effects can be accounted for by suitably modifying the parameters (e.g., mass) which appear in the non-interacting Hilbert space $\mathcal{H}$. This procedure is associated with what is called renormalization.

2. There exist bound states, e.g., the hydrogen atom.

3. Suppose we decide that we shall allow incoming states containing only photons and electron-positrons. Then, if the energies are sufficiently large, the outgoing states could certainly include π-mesons, proton-antiproton pairs, etc. The problem is that we do not have an exhaustive list of all "elementary" particles, and so we cannot write down the "final" $\mathcal{H}$. We are forced to proceed by a series of approximations. In certain situations, interactions which produce elementary particles not included in our $\mathcal{H}$ will not play an important role. Thus, we can closely describe the physical situation by using only one or two of the many interactions between elementary particles. Whenever we write down an $\mathcal{H}$ and an S, we have a physical theory with only a limited range of validity. The larger our $\mathcal{H}$, and the more interactions included, the larger is the domain of validity of our theory.

Despite these objections, we shall, as our starting point for the discussion of interactions, use the S-matrix approach.

27 The Hilbert Space of Interacting States

We have seen in Sect. 26 that scattering phenomena are completely described by a certain unitary operator S on a Hilbert space $\mathcal{H}$. We also remarked that, since $\mathcal{H}$ represents noninteracting states, and since the states of the actual physical system are influenced by interactions, we cannot interpret $\mathcal{H}$ as encompassing the states of our system. Is it possible, then, to construct a Hilbert space $\mathcal{L}$ which does represent the states of the interacting system? The answer is yes (at least, for scattering states), provided we accept a sufficiently loose interpretation for the word "construct."

What features would we expect for a Hilbert snare which is to represent the "interacting states of the system"? Firstly, comparing $\mathcal{L}$ and $\mathcal{H}$ in the distant past, we might expect to have an isomorphism $\Lambda_{\text{in}} : \mathcal{L} \to \mathcal{H}$ between $\mathcal{L}$ to $\mathcal{H}$. (See the example on p. 92.) (An *isomorphism* between two Hilbert spaces is a mapping from one to the other which is one-to-one, onto, linear, and norm-preserving. Clearly, any isomorphism has an inverse, which is itself an isomorphism.) Similarly, we would expect to have a second isomorphism $\Lambda_{\text{out}} : \mathcal{L} \to \mathcal{H}$. Finally, from the definition of the S-matrix, we would expect to have

$$S = \Lambda_{\text{out}} \Lambda_{\text{in}}^{-1} \tag{432}$$

Fix $\mathcal{H}$ and S. A triple, $(\mathcal{L}, \Lambda_{\text{in}}, \Lambda_{\text{out}})$ consisting of a Hilbert space $\mathcal{L}$ and two isomorphisms, Λ_{in} and Λ_{out}, from $\mathcal{L}$ to $\mathcal{H}$, subject to (432), will be called on *interaction space*. How many essentially different interaction spaces are there for a given $\mathcal{H}$, S? In fact, there is just one, in the following sense: Let $(\mathcal{L}', \Lambda'_{\text{in}}, \Lambda'_{\text{out}})$ and $(\mathcal{L}, \Lambda_{\text{in}}, \Lambda_{\text{out}})$ be two interaction spaces for $\mathcal{H}$, S. Then there exists a unique

isomorphism Ψ from $\mathcal{L}$ to $\mathcal{L}'$ such that

$$\Lambda_{\text{in}} = \Lambda'_{\text{in}} \cdot \Psi \qquad \Lambda_{\text{out}} = \Lambda'_{\text{out}} \cdot \Psi \tag{433}$$

(Proof: Evidently, we must choose $\Psi = {\Lambda'_{\text{in}}}^{-1}\Lambda_{\text{in}}$, which, by (432), is the same as $\Psi = {\Lambda'_{\text{out}}}^{-1}\Lambda_{\text{out}}$.) That is to say, the interaction space is unique up to isomorphism (which, of course, is as unique as we could expect it to be.)

All this looks rather pedagogical. After all, $\mathcal{L}$ is just another copy of $\mathcal{H}$, so why don't we just say so instead of speaking of "triples," etc.? The point is that the interaction space is more than just another cony of $\mathcal{H}$ — it also contains, as part of its structure, certain isomorphisms from $\mathcal{L}$ to $\mathcal{H}$. As a consequence, only certain portions of the (extensive) structure on $\mathcal{H}$ can be carried over, in a natural way, to $\mathcal{L}$. Examples of structure on $\mathcal{H}$ — which arise from the way in which $\mathcal{H}$ was constructed — are:

1. The total charge operator on $\mathcal{H}$.
2. The "total number of photons" operator on $\mathcal{H}$ (if, say, $\mathcal{H}$ happens to include photons).
3. The projection operator onto photon states (eliminating all other types of particles).
4. The operators which arise from the action of the restricted Poincaré group on $\mathcal{H}$.
5. The "number of baryons minus number of anti-baryons" operator.
6. The creation and annihilation operators of various particles in various states.
7. The field operators on $\mathcal{H}$.

The important point is that, in every case, the "additional structure" on $\mathcal{H}$ can be described by giving an operator on $\mathcal{H}$. The question of transferring structure from $\mathcal{H}$ to $\mathcal{L}$ reduces, therefore, to the following: under what conditions does an operator A on $\mathcal{H}$ define, in a natural way, an operator on $\mathcal{L}$? In fact, given an operator A on $\mathcal{H}$, there are two natural operators on $\mathcal{L}$, namely,

$$\Lambda_{\text{in}}^{-1} A \Lambda_{\text{in}} \qquad \Lambda_{\text{out}}^{-1} A \Lambda_{\text{out}} \tag{434}$$

In other words, we can carry A from $\mathcal{H}$ to $\mathcal{L}$ via either of the isomorphisms Λ_{in} or Λ_{out}. Which of (434) should we choose? There would he no choice if these two operators were equal. Thus, an operator A on $\mathcal{H}$ leads to a unique operator on the interaction space $\mathcal{L}$ provided A is such that the two operators (434) are equal. Using (432), this is equivalent to the condition

$$[S, A] = 0 \tag{435}$$

This, of course, is the result we expect. It is only properties of $\mathcal{H}$ which are invariant under the interaction (i.e., operators which commute with the S-matrix) which lead unambiguously to properties of the interaction states.

To summarize, structure on the interaction space is obtained from operators on $\mathcal{H}$ which commute with the S-matrix.

The operators which commute with S characterize what are called *conserved quantities*. They include charge, baryon number, lepton number, momentum, angular momentum, etc. Operators such as 2, 3, 6, and 7 above will not commute with S. They describe quantities which are not conserved in interactions.

28 Calculating the S-Matrix: An Example

We shall soon begin writing down formulae for the S-matrix. Unfortunately, these formula are rather complicated. They contain large numbers of terms, sums and integrals whose convergence is doubtful, and symbols whose precise meaning is rather obscure. We wish to avoid encountering all of these problems simultaneously. It is convenient, therefore, to first study a simpler example — a problem in which some of the features of the S-matrix formulae are exhibited, and in which some, but only some, of the difficulties are seen. We discuss such an example in the present section.

Let H he a fixed Hilbert space. Let $K(t)$ be a one-parameter family of bounded operators defined everywhere on H. That is, for each real number t, $K(t)$ is an operator on H. Suppose furthermore that $K(t) = 0$ unless t is in some finite interval. That is, suppose that there are numbers $t_i < t_f$ such that $K(t) = 0$ for $t \geq t_f$ and $t \leq t_i$. We are interested in studying curves in H, i.e., one-parameter families $\sigma(t)$ of elements of H, which satisfy the equation

$$-\frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}t}\sigma(t) = K(t)\sigma(t) \tag{436}$$

(Note: Derivatives and integrals of one-parameter families of elements of a Hilbert space, and operators on a Hilbert space, are defined by the usual limiting procedure.) Let $\sigma(t)$ satisfy (436). Then, since $K(t) = 0$ for $t \leq t_i$, $\sigma(t)$ is a constant element of H, σ_i, for $t \leq t_i$. Similarly, $\sigma(t) = \sigma_f$ for $t \geq t_f$. Clearly, a solution of (436) is completely and uniquely determined by σ_i, and σ_f is a linear function of σ_i. We write

$$\sigma_f = S\sigma_i \tag{437}$$

where S is some operator on H. The problem is to find an expression for S in terms of $K(t)$.

We first consider a special case in which the solution is easy. Suppose that all the $K(t)$'s commute with each other, i.e., $[K(t), K(t')] = 0$ for any t and t'. Then

$$\sigma(t) = \left[\exp\left(-\frac{i}{\hbar}\int_{t_i}^{t}\mathrm{d}\tau K(\tau)\right)\right]\sigma_i \tag{438}$$

is clearly a solution of (436). (Note: If $A(t)$ is a one-parameter (differentiable) family of operators on H, then $\frac{\mathrm{d}}{\mathrm{d}t}\exp A(t) = (\exp A(t))\frac{\mathrm{d}}{\mathrm{d}t}A(t)$ only when $A(t)$ and $\frac{\mathrm{d}}{\mathrm{d}t}A(t)$ commute.) Therefore,

$$\begin{aligned} S &= \exp\left(-\frac{i}{\hbar}\int_{t_i}^{t_f}\mathrm{d}\tau K(\tau)\right) \\ &= \mathbb{I} + \left(-\frac{i}{\hbar}\int_{t_i}^{t_f}\mathrm{d}\tau K(\tau)\right) + \frac{1}{2!}\left(-\frac{i}{\hbar}\int_{t_i}^{t_f}\mathrm{d}\tau K(\tau)\right)^2 \\ &\qquad + \frac{1}{3!}\left(-\frac{i}{\hbar}\int_{t_i}^{t_f}\mathrm{d}\tau K(\tau)\right)^3 + \cdots \end{aligned} \tag{439}$$

(The exponential of an operator is, of course, defined by the second equality in (439). We ignore, for the time being, questions of the convergence of such series.) Thus, when the $K(t)$'s commute, S is given by the relatively simple expression (439).

Now suppose that the $K(t)$'s do not commute. Integrating (436), we rewrite it as an integral equation:

$$\sigma(t) = \sigma_i - \frac{i}{\hbar}\int_{t_i}^{t}\mathrm{d}\tau K(\tau)\sigma(\tau) \tag{440}$$

We shall solve (440), at least formally, using a sequence of approximations. We begin with a trial solution, $\sigma_0(t)$. We substitute this $\sigma_0(t)$ into the right side of (440), and denote the result by $\sigma_1(t)$. We now take $\sigma_1(t)$ as our next trial solution. Substituting it into the right side of (440) to obtain $\sigma_2(t)$, etc. Thus, the general formula for passing from one trial solution to the next is

$$\sigma_{n+1}(t) = \sigma_i - \frac{i}{\hbar}\int_{t_i}^{t}\mathrm{d}\tau K(\tau)\sigma_n(\tau) \tag{441}$$

As our initial trial solution, we take $\sigma_0(t) = \sigma_i$, a constant. The hope is that the resulting $\sigma_n(t)$ will, as $n \to \infty$, converge, in a suitable sense, to a solution $\sigma(t)$ of (436). Using (441) successively, we have

$$\begin{aligned} \sigma_1(t) &= \sigma_i - \frac{i}{\hbar}\int_{t_i}^{t}\mathrm{d}\tau_1 K(\tau_1)\sigma_i \\ \sigma_2(t) &= \sigma_i - \frac{i}{\hbar}\int_{t_i}^{t}\mathrm{d}\tau_1 K(\tau_1)\sigma_i + \left(-\frac{i}{\hbar}\right)^2\int_{t_i}^{t}\mathrm{d}\tau_1\int_{t_i}^{\tau_1}\mathrm{d}\tau_2 K(\tau_1)K(\tau_2)\sigma_i \end{aligned} \tag{442}$$

or, more generally,

$$\sigma_n(t) = \left[\sum_{m=0}^{n} \left(-\frac{i}{\hbar} \right)^m \int_{t_i}^{t} \mathrm{d}\tau_1 \int_{t_i}^{\tau_1} \mathrm{d}\tau_2 \cdots \int_{t_i}^{\tau_{m-1}} \mathrm{d}\tau_m K(\tau_1) \cdots K(\tau_m) \right] \sigma_i \quad (443)$$

Thus, our formal limiting solution is

$$\sigma_n(t) = \left[\sum_{m=0}^{\infty} \left(-\frac{i}{\hbar} \right)^m \int_{t_i}^{t} \mathrm{d}\tau_1 \int_{t_i}^{\tau_1} \mathrm{d}\tau_2 \cdots \int_{t_i}^{\tau_{m-1}} \mathrm{d}\tau_m K(\tau_1) \cdots K(\tau_m) \right] \sigma_i \quad (444)$$

Indeed, if we substitute (444) into (436), and ignore questions of convergence of sums and the validity of interchanging the order of differentiation and summation, we obtain an identity. Thus, our formal expression for S is

$$S = \sum_{m=0}^{\infty} \left(-\frac{i}{\hbar} \right)^m \int_{t_i}^{t} \mathrm{d}\tau_1 \int_{t_i}^{\tau_1} \mathrm{d}\tau_2 \cdots \int_{t_i}^{\tau_{m-1}} \mathrm{d}\tau_m K(\tau_1) \cdots K(\tau_m) \quad (445)$$

It is convenient to recast (445) into a form which more closely resembles (439). The idea is to eliminate integrals whose limits of integration lie between t_i and t_f, i.e., to have all integrals be over the full range from t_i to t_f. Explicitly, the first few terms of (445) are

$$S = \mathbb{I} + \left(-\frac{i}{\hbar} \right) \int_{t_i}^{t_f} \mathrm{d}\tau_1 K(\tau_1) + \left(-\frac{i}{\hbar} \right)^2 \int_{t_i}^{t_f} \mathrm{d}\tau_1 \int_{\tau_1}^{t_f} \mathrm{d}\tau_2 K(\tau_1) K(\tau_2) + \left(-\frac{i}{\hbar} \right)^3 \int_{t_i}^{t_f} \mathrm{d}\tau_1 \int_{\tau_1}^{t_f} \mathrm{d}\tau_2 \int_{\tau_2}^{t_f} \mathrm{d}\tau_3 K(\tau_1) K(\tau_2) K(\tau_3) + \cdots \quad (446)$$

The first two terms on the right in (446) are already in the desired form. However, the third term on the right,

$$\left(-\frac{i}{\hbar} \right)^2 \int_{t_i}^{t_f} \mathrm{d}\tau_1 \int_{\tau_1}^{t_f} \mathrm{d}\tau_2 K(\tau_1) K(\tau_2) \quad (447)$$

is not. The region of integration in (447) is shown in the figure. The idea is to reverse the orders of the two integrations, while keeping the actual region over which the integration is performed — the shaded region in Figure 4 — unchanged. Thus, the expression (447) is equal to

$$\left(-\frac{i}{\hbar} \right)^2 \int_{t_i}^{t_f} \mathrm{d}\tau_2 \int_{\tau_2}^{t_f} \mathrm{d}\tau_1 K(\tau_1) K(\tau_2) \quad (448)$$

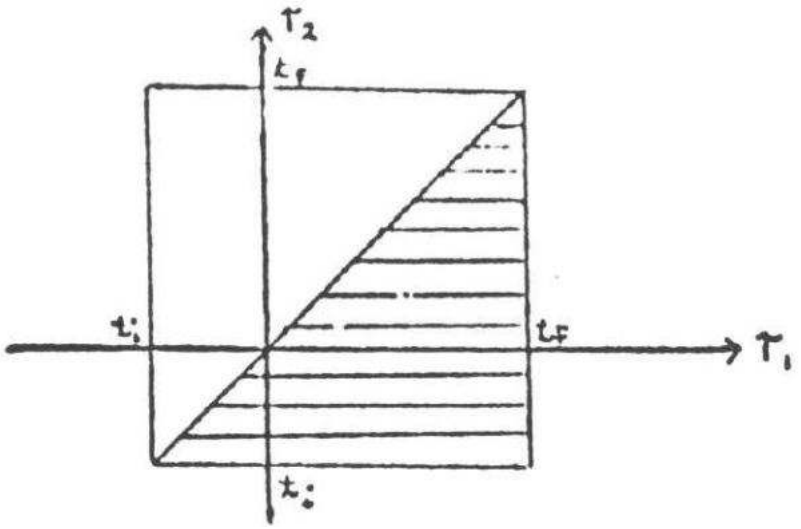


Figure 4: Region of integration in (447).

We next reverse the roles of the integration variables, τ_1 and τ_2, in (448) to obtain

$$\left(-\frac{i}{\hbar}\right)^2 \int_{t_i}^{t_f} \mathrm{d}\tau_1 \int_{\tau_1}^{t_f} \mathrm{d}\tau_2 K(\tau_2)K(\tau_1) \tag{449}$$

Finally, adding (447) and (449), we find that (447) is equal to

$$\frac{1}{2}\left(-\frac{i}{\hbar}\right)^2 \int_{t_i}^{t_f} \mathrm{d}\tau_1 \int_{t_i}^{t_f} \mathrm{d}\tau_2 T\left[K(\tau_1), K(\tau_2)\right] \tag{450}$$

where we have defined

$$T\left[K(\tau_1), K(\tau_2)\right] = \begin{cases} K(\tau_1)K(\tau_2) & \text{if } \tau_1 \geq \tau_2 \\ K(\tau_2)K(\tau_1) & \text{if } \tau_2 \geq \tau_1 \end{cases} \tag{451}$$

A similar procedure can be applied to each successive term in (446). The n^{th} term is equal to

$$\frac{1}{n!}\left(-\frac{i}{\hbar}\right)^n \int_{t_i}^{t_f} \mathrm{d}\tau_1 \cdots \int_{t_i}^{t_f} \mathrm{d}\tau_n T\left[K(\tau_1), \ldots, K(\tau_n)\right] \tag{452}$$

where $T[K(\tau_1), K(\tau_2), \ldots, K(\tau_n)]$ is defined to be the product of these operators, but arranged in the order in which the operator associated with the smallest

τ_m is placed on the right, the operator associated with the next-smallest τ_m is placed next, etc. This $T[K(\tau_1), K(\tau_2), \ldots, K(\tau_n)]$ is called the *time-ordered product*. Thus, our final formal expression for S is

$$S = \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\frac{i}{\hbar}\right)^n \int_{t_i}^{t_f} \mathrm{d}\tau_1 \int_{t_i}^{t_f} \mathrm{d}\tau_2 \cdots \int_{t_i}^{t_f} \mathrm{d}\tau_n T\left[K(\tau_1), \ldots, K(\tau_n)\right] \tag{453}$$

Note that, if all the $K(\tau)$'s commute, then the time-ordering is irrelevant, and (453) reduces to (439).

To summarize, the only modification of (439) required when the operators do not commute is that the products of operators in the multiple integrals must be time-ordered.

Finally, note that if all the $K(\tau)$'s are Hermitian, then, from (436),

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left(\sigma(t), \sigma(t)\right) &= \left(\frac{\mathrm{d}}{\mathrm{d}t}\sigma, \sigma\right) + \left(\sigma, \frac{\mathrm{d}}{\mathrm{d}t}\sigma\right) \\ &= \left(-\frac{i}{\hbar}K\sigma, \sigma\right) + \left(\sigma, -\frac{i}{\hbar}K\sigma\right) \\ &= \frac{i}{\hbar}\left[(K\sigma, \sigma) - (\sigma, K\sigma)\right] = 0 \end{aligned} \tag{454}$$

Hence, S must be a unitary operator. Unitarity is obvious in (439) — rather less so in (453).

Our formulae for the S-matrix in field theory will also involve infinite series of multiple integrals of time-ordered products of Hermitian operators.

29 The Formula for the S-Matrix

In this section we shall write down the formula for the S-matrix in terms of a certain (as yet unspecified) operator field on Minkowski space. While we shall in no sense "derive" that our expression for S is correct, it will be possible, at least, to show that the formula is a reasonable guess. We rely heavily on the discussion in Sect. 28.

Consider Eqn. (453). We wish to write down an analogous formula for S in quantum field theory. Since, first of all, S is to be an operator on the non-interacting Hilbert space we must take the K's to be operators on $\mathcal{H}$. Secondly, S should be unitary: we therefore take the K's to de, Hermitian. By what should we replace the interaction variables - the τ's in Eqn. (453)? If we think of τ in Sect. 28 as representing a "time", then a natural replacement would be position x in Minkowski space-time. The integrals in (453) would then extend over all of Minkowski space. (Note: This is the reason why it was convenient, in Sect. 28, to obtain an expression in which the integrals extended over the entire t-range from t_i to t_f.) Thus, we are led to consider the interaction is described by a certain Hermitian operator field, $\underline{K}(x)$, which depends on position x in

Minkowski space, and acts on $\mathcal{H}$. The S-matrix will then be given by the formula

$$S = \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\frac{i}{\hbar}\right)^n \int \mathrm{d}V_1 \cdots \int \mathrm{d}V_n T\left[\underline{K}(x_1), \ldots, \underline{K}(x_n)\right] \qquad (455)$$

where $x_1, x_2, \ldots$ represent points in Minkowski space, $\mathrm{d}V_1, \mathrm{d}V_2, \ldots$ the corresponding volume elements in Minkowski space, and all integrals are over all of Minkowski space. Note that Eqn. (453) already suffers from one difficulty: the question of the convergence of the infinite series. In writing (455) we have retained that difficulty, and, in fact , introduced a second one: the question of the convergence of the integrals. (The integrals in (453) are all over compact regions.)

In fact, there is a second problem with (455) which was not encountered in (453), namely, the question of what the time-ordering operator T is to mean in (455). In (453), T means that the $K(\tau)$ operators are to be placed in the order of decreasing τ-values. Unfortunately, in the passage from a "time" τ to position x in Minkowski space-time, the natural ordering is destroyed. There is, however, one case in which points in Minkowski space can be ordered: we agree that x_2 exceeds x_1, $x_2 > x_1$, if $x_2 - x_1$ (i.e., the position vector of x_2 relative to x_1) is future-directed and timelike or null. Hence, for the region of integration in (455) for which all the $x_1, \ldots, x_n$ can be ordered in this way, T has a well-defined meaning, and hence the integral makes sense. (More explicitly, the τ's are totally ordered, while points in (time-oriented) Minkowski space are only partially ordered.) Clearly, there is no Poincaré-invariant way to "time-order" $\underline{K}(x_1)$ and $\underline{K}(x_2)$ when $x_2 - x_1$ is spacelike. How, then, are we to give a meaning to (455)? One way of doing this would be through an additional condition on the $\underline{K}(x)$. We simply assume that the ordering of $\underline{K}(x_1)$ and $\underline{K}(x_2)$ is irrelevant when $x_2 - x_1$ is spacelike. That is to say, a natural way of forcing a meaning for (455) is to assume that the $\underline{K}(x)$ have the property

$$[\underline{K}(x), \underline{K}(x')] = 0 \qquad \text{for } x - x' \text{ spacelike} \qquad (456)$$

We include (456) as a requirement on our $\underline{K}(x)$. Thus, if $x_1, x_2, \ldots, x_n$ are points in Minkowski space, we define $T[\underline{K}(x_1), \underline{K}(x_2), \ldots, \underline{K}(x_n)]$ to be the product of these operators, placed in an order such that, if $x_i - x_j$ is timelike or null and future-directed, then $\underline{K}(x_i)$ appears before $\underline{K}(x_j)$ in the product. Clearly, there always exists at least one ordering having this property. Furthermore, (456) implies that all such orderings yield the same operator. Thus, $T[\underline{K}(x_1), \ldots, \underline{K}(x_n)]$ is a well-defined operator.

To summarize, the interaction is to be described by giving a Hermitian operator field $\underline{K}(x)$ on $\mathcal{H}$ which satisfies (456). The S-matrix is to be expressed in terms of $\underline{K}(x)$ by (455). The formal expression (455) is unsatisfactory insofar as we have investigated the convergence of neither the infinite series nor the integrals themselves.

The standard textbooks give a more detailed, but, I feel, no more satisfactory, argument for (455). One chooses a time-coordinate t in Minkowski space,

and writes

$$\underline{H}(t) = \int\limits_{t=\text{const.}} \underline{K}(x) \tag{457}$$

for the "Hamiltonian". One then imagines a "state vector" in the Hilbert space $\mathcal{H}$, $\sigma(t)$, which depends on the time-coordinate. One writes the "Schrödinger equation",

$$-\frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}t}\sigma(t) = \underline{H}(t)\sigma(t) \tag{458}$$

The argument of Sect. 28 then yields (455), where T refers to the ordering induced by the time-coordinate t. Finally Poincaré invariance (i.e., the condition that S be independent of our original choice of t) requires (456). This argument assumes, implicitly, that the "states during the interaction" are described by $\mathcal{H}$. We have deviated only slightly from this conventional argument. We isolated the simplest and clearest part in Sect. 28, and guessed the rest.

30 Dimensions

With each of the various fields and operators we have introduced, there is associated a corresponding physical dimension. We shall determine these dimensions in the present section.

Recall that we have set the speed of light equal to one, so length and time have the same units. (E.g., we measure distance in light-seconds.) We may therefore take as our fundamental units a mass (m) and a time (t). Then Planck's constant h has dimensions mt. We assign to position vectors in Minkowski space dimensions t, so the derivative in Minkowski space has dimensions t^{-1}, and the wave operator $\Box$ dimensions t^{-2}. (Raising, lowering, and contracting indices does not affect dimensions.) The quantity μ which appears in the Klein-Gordon and Dirac equations therefore has dimensions t^{-1}. Position vectors in momentum space have dimensions t^{-1}. Finally, the volume element on the mass shell has dimensions t^{-1} (see (14).)

The rule for determining the dimensions to be associated with a classical field is the following: consider an element of the Hilbert space which has norm unity, and work back to determine the dimensions of the corresponding field. Consider first the (real or complex) Klein-Gordon case. A unit vector in the Hilbert space is represented by a function $\phi(k)$ on M_μ which satisfies

$$\frac{1}{\hbar}\int\limits_{M_\mu} \phi(k)\overline{\phi(k)}\,\mathrm{d}V_\mu = 1 \tag{459}$$

Therefore, $\phi(k)$ has dimensions $m^{1/2}t^{3/2}$. But

$$\phi(x) = \int \phi(k)e^{ik_b x^b}\,\mathrm{d}V \tag{460}$$

and so the Klein-Gordon field has dimensions $m^{1/2}t^{-1/2}$. For the Dirac case, Eqn. (388) implies that $(\xi^A(k), \eta_{A'}(k)$ has dimensions $t^{1/2}$. Then (309) and

(310) imply that $(\xi^A(k), \eta_{A'}(k)$ has dimensions $t^{3/2}$. The neutrino fields have the same dimensions. Finally, for the Maxwell case, (382) implies that $A_a(k)$ has dimensions $m^{1/2}t^{3/2}$, whence, from (178), $A_a(x)$ has dimensions $m^{1/2}t^{-1/2}$. (In the Maxwell case, one has a simple independent check on the dimensions. The dimensions above for the vector potential imply that electric and magnetic fields have dimensions $m^{1/2}t^{3/2}$ which agrees, of course, with the dimensions from classical electrodynamics.)

The other quantities whose dimensions are of particular interest are the field operators. To assign dimensions to these operators, we must decide what dimensions are to be associated with the test fields. Recall that the role of the test fields is to smear out the (undefined) operators associated with points of Minkowski space, e.g.,

$$\underline{\phi}(f) = \int f(x)\underline{\phi}(x)\,\mathrm{d}V \tag{461}$$

Thus, we can think of a test field as a "smearing density." We therefore take all test fields to have dimensions t^{-4}. (That is, we require that $\underline{\phi}(f)$ and the (undefined) $\underline{\phi}(x)$ have the same dimensions.) With this convention, the determination of the dimensions of the field operators is straightforward. Consider first the Klein-Gordon case. If $f(x)$ is a test field (dimensions t^{-4}), then, from

$$f(x) = \int f(k)e^{ik_b x^b}\,\mathrm{d}V \tag{462}$$

$f(k)$ is dimensionless. But a dimensionless element of our Hilbert space, (459), defines a $\phi(k)$ with dimensions $m^{1/2}t^{3/2}$. Therefore, the element of our Hilbert space associated with this $f(k)$, $\sigma(f)$, has dimensions $m^{-1/2}t^{-3/2}$. Thus, the creation and annihilation operators have dimensions $m^{-1/2}t^{-3/2}$. But

$$\underline{\phi}(f) = \hbar C(\sigma(f)) + \hbar A(\sigma(f)) \tag{463}$$

(say, for real Klein-Gordon fields), and so the field operators, $\underline{\phi}(f)$, have dimensions $m^{1/2}t^{1/2}$. In the Dirac case, the test fields, $(f^A(x), f_{A'}(x))$ have dimensions t^{-4}, whence $f^A(k)$ is dimensionless. The corresponding pair of functions on the mass shell, (389), therefore has dimensions t^{-1}. Hence, the corresponding elements of our Hilbert space, $\sigma(f^A, f_{A'})$, have dimensions $t^{-3/2}$. This, then, is the dimensions of the creation and annihilation operators. Finally, from (390), the field operators in the Dirac case have dimensions $t^{-3/2}$. Similarly, in the Maxwell case, $f^a(x)$ has dimensions t^{-4}, $f^a(k)$ dimensionless, $\sigma(f^a)$ dimensions $m^{-1/2}t^{-3/2}$. The creation and annihilation operators therefore have dimensions $m^{-1/2}t^{-3/2}$, and so by (383), the Maxwell field operators have dimensions $m^{1/2}t^{-1/2}$.

Note that, in every case, the classical fields and the corresponding field operators have the same dimensions.

31 Charge Reversal

An important tool, both for constructing and analyzing interactions, is the discrete symmetries — charge, parity, and time reversal. We begin with charge reversal.

Let $\mathcal{H}$ be a non-interacting Hilbert space, and suppose we are given an operator S, the S-matrix, on $\mathcal{H}$. What does it mean to say that "the interaction described by S is invariant under charge reversal?" Roughly speaking, this means that if we replace any in-state, $\sigma \in \mathcal{H}$, by "the same state, but with all particles replaced by their antiparticles," then the corresponding out-states are again "the same, but with particles replaced by antiparticles." Thus, we are led to try to give a meaning to the notion "the same state (in $\mathcal{H}$), but with particles replaced by antiparticles." Let us suppose that this mapping from non-interacting states to non-interacting states is accomplished by come operator C on $\mathcal{H}$, so that $C\sigma$ represents the same state as σ, but with particles replaced by antiparticles. Then the statement that the interaction (described by S) is invariant under charge reversal reduces to the condition

$$SC\sigma = CS\sigma \tag{464}$$

for any $\sigma \in \mathcal{H}$. In other words, invariance under charge reversal is expressed mathematically by the condition that S and C (both operators on $\mathcal{H}$) commute.

In general, there will be a number of different operators C which could be interpreted as effecting the replacement of particles by antiparticles. There is no obvious, unambiguous way of translating this physical notion into a mathematical operator. We shall therefore proceed as follows. We first write down a list of properties which reflect the intuitive idea of "replacing particles by antiparticles, but not otherwise changing the state." In general, there will he a moderately large class of C's which satisfy these criteria. Then, for each interaction, we look for an operator C which satisfies our criteria and which, in addition, commutes with the S-matrix. If such a C exists, we say that our interaction is invariant under charge-reversal. The point is that *any* operator which commutes with the S-matrix is valuable. We regard the words "charge reversal" as merely suggesting a particularly fertile area in which such operators might be found. This philosophy is important:

i) there is no natural, *a priori* charge-reversal operator;

ii) one sets up a class of possible charge-reversal operators, and then selects from this class depending on what the interaction is,

iii) if no operator in this class commutes with the S-matrix, there is little point in considering charge-reversal for that interaction.

(The third point is somewhat over-stated, and will be modified slightly later.)

The first condition on C is that it should not mix up particles of different types. That is to say, C should commute with the total number operators on each of the Fock spaces which make up $\mathcal{H}$. Therefore, C can be considered

as an operator on each of the Fock spaces separately. The assumption that C commutes with the total number operators implies, furthermore, that C can be decomposed into an operator on the one-particle Hilbert space, an operator on the two-particle Hilbert space, etc. We next assume that the action of C on the many-particle states can be obtained from the action on the one-particle states as follows. Let H be one of our one-particle Hilbert spaces (e.g., H_{RKG}, H_{CKG}, H_M, H_D). Then an element of the corresponding Fock space consists of a string

$$(\xi, \xi^\alpha, \xi^{\alpha\beta}, \ldots) \tag{465}$$

of tensors over H. The operator C on the one-particle Hilbert space H can be written, in the index notation, as $C^\alpha{}_\beta$: the result of applying $C^\alpha{}_\beta$ to an element ξ^β of H is written $C^\alpha{}_\beta \xi^\beta$. We assume that the action of C on the element (465) of Fock space is

$$(\xi, C^\alpha{}_\beta \xi^\beta, C^\alpha{}_\mu C^\beta{}_\nu \xi^{\mu\nu}, \ldots) \tag{466}$$

This is a quite reasonable assumption: if we know what charge-reversal means on a one-particle state, we assume that, for a two-particle state, the effect of charge reversal is to "apply charge-reversal to each of the particles individually."

Thus, we are led to distinguish a class of charge-reversal operators on each of our one-particle Hilbert spaces, H_{RKG}, H_M, etc.

It is convenient to introduce some definitions. A mapping T from a Hilbert space H to itself is said to be *antilinear* if

$$T(\alpha\sigma + \tau) = \bar{\alpha} T(\sigma) + T(\tau) \tag{467}$$

for any $\sigma, \tau \in H$, $\alpha \in \mathbb{C}$. (Alternatively, T could be considered as a linear mapping from H to $\bar{H}$.) We shall sometimes refer to an antilinear mapping as an antilinear operator. The word "operator" alone means "linear operator." A linear or antilinear operator T is said to be *norm-preserving* if

$$\|T\sigma\| = \|\sigma\| \tag{468}$$

for every $\sigma \in H$. Eqn. (468) immediately implies that, for any $\sigma, \tau \in H$,

$$(T\sigma, T\tau) = (\sigma, \tau) \tag{469}$$

$$\text{or} \qquad (T\sigma, T\tau) = (\tau, \sigma) \tag{470}$$

according as T is linear or antilinear, respectively. As we have remarked, a linear, norm-preserving operator is called unitary. An antilinear, norm-preserving operator is said to be *antiunitary*.

Let H be one of the Hilbert spaces H_{RKG}, H_{CKG}, H_M, or H_D. A linear or antilinear operator C on H will be called a *charge-reversal operator* if

1. C is norm-preserving.

2. C commutes with all the unitary operators on H which arise from the action of the restricted Poincaré group on H.

3. C, applied to a positive- (resp., negative-) frequency element of H, yields a negative- (resp., positive-) frequency element.

These conditions are to reflect the intuitive that "the state is changed only in that particles are replaced by antiparticles." Conditions 1 and 3 are clearly reasonable. (The passage from positive-frequency to negative-frequency is the passage from particles to anti-particles.) Condition 2 ensures that quantities such as the locations and momenta of particles are unchanged under C. Note that, if C is a charge-reversal operator, and α is a complex number with $|\alpha| = 1$, then αC is also a charge-reversal operator. (One could, conceivably, impose the further condition $C^2 = 1$. We shall not do so.) Note that Condition 1 implies that C is also norm-preserving on the non-interacting Hilbert space $\mathcal{H}$.

Before discussing examples of charge-reversal operators, we establish the following result: C must be linear rather than antilinear. Let r^a be a constant, unit, future-directed, timelike vector field in Minkowski space. Then the energy operator associated with r^a is

$$\underline{E} = \frac{i}{\hbar} r^a \underline{L}_a \tag{471}$$

where $r^a \underline{L}_a$ is the operator which comes from the unitary transformation associated with the Poincaré transformation (a translation) generated by r^a. It is essential, in (471), that the $i/\hbar$ appear explicitly, so that r^a is simply the first-order difference between a unitary operator and the identity operator. We assume that C is antiunitary, and obtain a contradiction. For each of our Hilbert spaces, the expectation value of the energy $\underline{E}$ in any state (and, in particular, in the state $C^{-1}\sigma$, for $\sigma \in H$) is non-negative:

$$(C^{-1}\sigma, \underline{E} C^{-1}\sigma) \geq 0 \tag{472}$$

But, from (470), this implies

$$(C\underline{E}C^{-1}\sigma, \sigma) \geq 0 \tag{473}$$

Write

$$C\underline{E}C^{-1} = \left(C \frac{i}{\hbar} C^{-1} \right) (C r^a \underline{L}_a C^{-1}) \tag{474}$$

Condition 2 above implies that $C(r^a \underline{L}_a)C^{-1}$. The assumption that C is antiunitary implies $C(i/\hbar)C^{-1} = -i/\hbar$. Thus, we have

$$(C^{-1}\sigma, \underline{E} C^{-1}\sigma) = -(\underline{E}\sigma, \sigma) \tag{475}$$

But this is a contradiction, for the left side is non-negative and the right side non-positive. Therefore, C cannot be anti-linear.

The question of the uniqueness of charge-reversal operators is settled by the following fact: let O be a bounded operator defined everywhere on H (one of our four one-particle Hilbert spaces). Suppose that O commutes with the unitary operators which arise from the action of the restricted Poincaré group

and, furthermore, that O takes positive- (resp., negative-) frequency to positive- (resp., negative-) frequency states. Then O is a multiple of the identity. I know of no simple proof. (The statement is essentially an infinite-dimensional generalization of Schur's Lemma (p. 58).) Now suppose that C and C' are charge-reversal operators. Then $C'C^{-1}$ satisfies the conditions above, and hence must be a multiple of the identity. It follows that $C' = \alpha C$, where α is a complex number with $|\alpha| = 1$. Thus, having found one charge-reversal operator on our Hilbert space H, we have found them all.

We begin with H_{RKG}. In this case, the only positive-frequency or negative-frequency solution is the zero solution. (That is to say, H_{RKG} describes particles which are identical with their antiparticles.) Hence, Condition 3 is empty. Therefore, the identity is a possible charge-reversal operator. We conclude that the most general charge-reversal operator on H_{RKG} is $\alpha\mathbb{I}$, with $|\alpha| = 1$.

For H_{CKG}, one charge-reversal operator is given by

$$\phi(x) \to \bar{\phi}(x) \tag{476}$$

or, in momentum space, by

$$\phi(k) \to \bar{\phi}(-k) \tag{477}$$

That this operator is unitary rather than antiunitary follows from our complex vector-space structure on H_{CKG} (see Sect. 12). Thus, the most general charge-reversal operator on H_{CKG} is given by

$$\phi(x) \to \alpha\bar{\phi}(x) \tag{478}$$

with $|\alpha| = 1$.

The most general charge-reversal operator on H_M is $\alpha\mathbb{I}$, with $|\alpha| = 1$.

The most general charge-reversal operator on H_D is

$$(\xi^A, \eta_{A'}) \to \alpha(\bar{\eta}^A, \bar{\xi}_{A'}) \tag{479}$$

with $|\alpha| = 1$.

32 Parity and Time Reversal

The basic idea of the remaining two discrete symmetries — parity and time reversal — is essentially the same as that for charge reversal. One is concerned primarily with finding operators which commute with the S-matrix, and operators which can be interpreted as representing parity and time reversal are particularly good candidates.

We begin with some remarks concerning the Poincaré group. Recall that the restricted Poincaré group, $\mathcal{RP}$, is a connected, 10-dimensional Lie group. This $\mathcal{RP}$ is a normal subgroup of another 10-dimensional Lie group, the full Poincaré group $\mathcal{P}$. However, $\mathcal{P}$ is not connected; it has four connected components. These components consist of Poincaré transformations which reverse neither time nor space orientation ($\mathcal{RP}$), time but not space orientation, space but not time

orientation, and both time and space orientation. The quotient group, $\mathcal{P}/\mathcal{RP}$, is isomorphic with the group $\mathbb{Z}_2 \times \mathbb{Z}_2$. ($\mathbb{Z}_2$ is the additive group of integers mod 2.)

The situation is slightly different for the boson and fermion cases. We begin with the boson case. Let H be one of the Hilbert spaces H_{RKG}, H_{CKG}, or H_M. Then, as we have seen (Sect. 16), H defines a representation of $\mathcal{RP}$. That is to say, with each $P \in \mathcal{RP}$ there is associated a unitary operator U_P on H, where these U_P's satisfy:

$$\begin{aligned} U_P U_{P'} &= U_{PP'} \\ U_e &= \mathbb{I} \end{aligned} \tag{480}$$

The problem of obtaining parity-reversal and time-reversal operators can be stated as follows: we wish to extend this representation from $\mathcal{RP}$ to $\mathcal{P}$. That is to say, we wish to find, for each $P \in \mathcal{P}$, a (unitary or antiunitary) operator U_P, subject to (480) and to the condition that, for $P \in \mathcal{RP}$, this representation reduces to the given representation (Sect. 16) of it $\mathcal{RP}$. It is necessary to admit both unitary and antiunitary operators for, as we shall see shortly, it is otherwise impossible to find any extension of our original representation of $\mathcal{RP}$.

There is an important difference between charge reversal on the one hand and parity and time reversal on the other. In the case of charge reversal, one settles eventually on a single unitary charge-reversal operator C. There is, however, no one natural "parity-reversal operator P" or "time-reversal operator T". There is, instead, a 10-dimensional manifold's worth of such operators, namely, the operators associated with the appropriate component of the Poincaré group.

Suppose now that we have a representation of $\mathcal{P}$ as described above. Let P and Q be in the same component of $\mathcal{P}$, so $PQ \in \mathcal{RP}$. Now $U_P U_Q = U_{PQ}$. But, since $PQ \in \mathcal{RP}$, U_{PQ} is unitary rather than antiunitary. We conclude that either both U_P and U_Q are unitary, or else both are antiunitary. (The product of two unitary operators, or two antiunitary operators, is unitary; the product of a unitary and an antiunitary operator is antiunitary.) We conclude that all the operators associated with the Poincaré transformations in a given component of $\mathcal{P}$ are the same (all unitary or all antiunitary.)

In fact, it follows from an argument similar to that used for charge-reversal that U_P is antiunitary if and only if the Poincaré transformation P reverses time. From the remarks above, it suffices to show that, for some P which reverses parity but not time, U_P is unitary, and that, for some P which reverses time but not parity, U_P is antiunitary. Let r^a be a constant, unit, future-directed timelike vector field, and consider the energy operator $\underline{\underline{E}}$ given by (471). All expectation values of $\underline{\underline{E}}$ are non-negative. Fix an origin O, and let t^a be a unit, future-directed timelike vector at O. Let P denote the Poincaré transformation which sends

$$x^a \to x^a + 2t^a(x^b t_b) \tag{481}$$

where x^a is the position vector relative to O. Evidently, this P reverses time orientation but not space orientation. From the commutativity properties of

the Poincaré group,

$$U_P^{-1}\underline{E}U_P = \left(U_P^{-1}\frac{i}{\hbar}U_P\right) r'^a \underline{L}_a \tag{482}$$

where

$$r'^a = r^a + 2t^a(r^b t_b) \tag{483}$$

Since r'^a is a past-directed timelike vector, the positivity of $\underline{E}$ implies, evidently, that U_P must be antiunitary. Similarly, let Q be the Poincaré transformation

$$x^a \to -x^a - 2t^a(x^b t_b) \tag{484}$$

So Q reverses spatial orientation but not time orientation. Then

$$U_Q^{-1}\underline{E}U_Q = \left(U_Q^{-1}\frac{i}{\hbar}U_Q\right) r'^a \underline{L}_a \tag{485}$$

where

$$r'^a = -r^a - 2t^a(r^b t_b) \tag{486}$$

Clearly, in this case r'^a is a future-directed timelike vector, hence U_Q must be unitary.

We next consider uniqueness. Let U_P and U'_P be two extensions of our representation of $\mathcal{RP}$, so $U_P = U'_P$ for $P \in \mathcal{RP}$. Let Q be a Poincaré transformation which reverses, say, temporal orientation but not spatial orientation. Consider the unitary operator

$$A = U'_Q U_Q^{-1} \tag{487}$$

We first how that this A depends only on the component of $\mathcal{P}$ in which Q lies. Let W be another Poincare transformation which lies in the same component of $\mathcal{P}$ as Q. Then $W = QR$ for some $R \in \mathcal{RP}$. Hence,

$$U'_W U_W^{-1} = U'_{QR}U_{QR}^{-1} = (U'_Q U'_R)(U_R^{-1}U_Q^{-1}) = U'_Q U_Q^{-1} \tag{488}$$

where we have used the fact that $U_R = U'_R$. We next show that, for $P \in \mathcal{RP}$, U_P commutes with A. Indeed, since $PQ = QV$ for some $V \in \mathcal{RP}$, we have

$$\begin{aligned} U_P A U_P^{-1} &= U_P U'_Q U_Q^{-1} U_P^{-1} = U'_{PQ}U_{PQ}^{-1} \\ &= U'_{QV}U_{QV}^{-1} = (U'_Q U_V)(U_V^{-1}U_Q^{-1}) = U'_Q U_Q^{-1} = A \end{aligned} \tag{489}$$

These properties do not yet suffice, however, to show that A is a multiple of the identity (see p. 106). We must impose an additional condition which ensures that A does not mix up particles and antiparticles. However, it is reasonable, on physical grounds, to make the following additional assumption: U_P reverses the roles of particles and antiparticles if and only if P reverses time directions. Under this assumption, A must be a multiple of the identity, whence $U_Q = \alpha U'_Q$, where α is some complex number (the same for every on Poincaré transformation in the same component as Q). However, since $QQ \in \mathcal{RP}$, we have

$$U'_Q U'_Q = U'_{QQ} = U_{QQ} = U_Q U_Q \tag{490}$$

whence $\alpha = \pm 1$.

To summarize, we are interested in extending a given representation from $\mathcal{RP}$ to $\mathcal{P}$ in such a way that U_P reverses the role of particles and antiparticles when and only when P reverses temporal orientation. Every such extension has the property that U_P is antiunitary when and only when P reverses temporal orientation. The extension of the representation is unique except for the following possibilities: affix a minus sign to U_P whenever P reverses spatial parity, affix a minus sign to U_P whenever reverses temporal orientation, or affix a minus sign to U_P whenever P reverses spatial or temporal orientation, but not both. Thus, from one extension of the representation, it is easy to write down them all.

Finally, we write down a representation of $\mathcal{P}$ for each of our Hilbert spaces H_{RKG}, H_{CKG}, and H_M. Let P be a Poincaré transformation, and write Px for the point of Minkowski space to which P sends the point x. For the real and complex Klein-Gordon cases, the action of U_P on an element $\phi(x)$ of our Hilbert space is as follows:

$$\phi(x) \to \phi(Px) \tag{491}$$

This action clearly defines a representation of $\mathcal{P}$. For the Maxwell case, note that P is a smooth mapping from Minkowski space to Minkowski space, and hence P sends any vector field on Minkowski space to another vector field. This action defines U_P on H_M. For example, the two Poincaré transformations (481) and (484) have the following actions on the vector potential $A_a(x)$:

$$\begin{aligned} A_a(x) &\to A_a(Px) + 2t_a(t^b A_b(Px)) \\ A_a(x) &\to -A_a(Qx) - 2t_a(t^b A_b(Qx)) \end{aligned} \tag{492}$$

The situation with regard to H_D differs in some important respects from that above. The fundamental difference is that H_D does not define a representation of $\mathcal{RP}$. Instead, the operative group is what is usually called "inhomogeneous $SL(2,\mathbb{C})$": the (double) covering group of $\mathcal{RP}$. ($SL(2,\mathbb{C})$ is the (double) covering group of the restricted Lorentz group.) We shall denote this connected, 10-dimensional Lie group by $\mathcal{RS}$. If $\mathcal{RS}$ is to replace $\mathcal{RP}$, what group should replace $\mathcal{P}$? One could, in fact, introduce such a group, and attempt to extend to it our representation of $\mathcal{RS}$. It is simpler, however, to proceed in a slightly different way.

Let $P \in \mathcal{RP}$. Associated with P there are precisely two elements of $\mathcal{RS}$. The corresponding pair of unitary operators on H_D differ only in sign. Thus, we can regard H_D as a double-valued representation of $\mathcal{RP}$: with each $P \in \mathcal{RP}$, there is associated two unitary operators on H_D, these operators differing only in sign. These operators satisfy (480), modulo sign. The question in the fermion case is therefore the following: Can this double-valued representation of $\mathcal{RP}$ be extended to a double-valued representation of $\mathcal{P}$? The argument given earlier shows that the operators associated with $P \in \mathcal{P}$ are antiunitary if and only if P reverses time-orientation. The uniqueness situation is essentially the same, by the same argument.

There remains, therefore, only the task of specifying what the operators are for $P \notin \mathcal{RP}$. Since the (double-valued) action of $\mathcal{RP}$ on H_D is known, we need only specify U_P for one time-reversing P and one parity-reversing P. The action is as follows. For the Poincaré transformation (481),

$$(\xi^A(x), \eta_{A'}(x)) \to \pm\frac{i}{\sqrt{2}}(t^{AA'}\eta_{A'}(Px), t_{AA'}\xi^A(Px)) \tag{493}$$

and for the Poincaré transformation (484),

$$(\xi^A(x), \eta_{A'}(x)) \to \pm\frac{1}{\sqrt{2}}(t^{AA'}\eta_{A'}(Qx), t_{AA'}\xi^A(Qx)) \tag{494}$$

33 Extending Operators to Tensor Products and Direct Sums

We have now introduced a large number of operators — some on the one-particle Hilbert spaces and some of the Fock spacs. In order that such operators can be discussed relative to the S-matrix, however, their action must be defined on the non-interacting Hilbert space $\mathcal{H}$. Since $\mathcal{H}$ arises from two constructions — the tensor product and direct sum of Hilbert spaces — we are led to the problem of extending the action of operators through these two constructions.

We begin with the direct sum. Let $H_1, H_2, \ldots$ be a sequence of Hilbert spaces. Then the direct sum of this sequence, $H = H_1 \oplus H_2 \oplus H_3 \oplus \ldots$ is the Hilbert space consisting of sequences

$$\sigma = (\sigma_1, \sigma_2, \sigma_3, \ldots) \tag{495}$$

with $\sigma_i \in H_i$ for which the sum

$$\|\sigma\|^2 = \|\sigma_1\|^2 + \|\sigma_2\|^2 + \cdots \tag{496}$$

which defines the norm, converges. Now let $O_1, O_2, \ldots$ be a sequence of operators (O_i on H_i) which are either all linear or all antilinear. We then define an operator O on H as follows:

$$O\sigma = (O_1\sigma_1, O_2\sigma_2, O_3\sigma_3, \ldots) \tag{497}$$

Clearly, O is linear (resp., antilinear) provided the O_i are linear (resp., antilinear.) Note, furthermore, that if all the O_i are norm-preserving, so is O; if all the O_i are Hermitian, so is O; if all the O_i are projection operators, so is O. Of course, not every operator on H can be expressed in the form (497).

The tensor product is next. Let $H^\alpha, H^\beta, \ldots, H^\delta$ be a finite sequence of Hilbert spaces. Then the tensor product of this sequence, $H = H^\alpha \otimes H^\beta \otimes \cdots \otimes H^\delta$ is the Hilbert space obtained as the completion of the inner-product space consisting of all formal expressions of the form:

$$\xi^{\alpha\beta\cdots\delta} = \sigma^\alpha\tau^\beta\cdots\mu^\delta + \cdots + \lambda^\alpha\rho^\beta\cdots\nu^\delta \tag{498}$$

(The index indicates the Hilbert space to which a vector belongs.) Now let ${O^{\alpha'}}_{\alpha}, {O^{\beta'}}_{\beta}, \ldots, {O^{\delta'}}_{\delta}$ be a sequence of linear operators on $H^{\alpha}, H^{\beta}, \ldots, H^{\delta}$, respectively. We then can define an operator O on H by:

$$O(\xi^{\alpha\beta\cdots\delta}) = ({O^{\alpha'}}_{\alpha}\sigma^{\alpha})({O^{\beta'}}_{\beta}\tau^{\beta})\cdots({O^{\delta'}}_{\delta}\mu^{\delta}) + \cdots + ({O^{\alpha'}}_{\alpha}\lambda^{\alpha})({O^{\beta'}}_{\beta}\rho^{\beta})\cdots({O^{\delta'}}_{\delta}\nu^{\delta}) \quad (499)$$

If all the O's are unitary, so is O; if all the O's are Hermitian, so is O; if all the O's are projection operators, so is O. Now suppose that the given O-operators are antilinear rather than linear. We thus have linear mappings from $\bar{H}_{\alpha}$ to H^{α}, from $\bar{H}_{\beta}$ to H^{β}, etc. Thus, in the index notation, the O's would be written $O^{\alpha'\alpha}, O^{\beta'\beta}, \ldots, O^{\delta'\delta}$. In this case, we define the corresponding operator O on H by

$$O(\xi^{\alpha\beta\cdots\delta}) = (O^{\alpha'\alpha}\bar{\sigma}_{\alpha})(O^{\beta'\beta}\bar{\tau}_{\beta})\cdots(O^{\delta'\delta}\bar{\mu}_{\delta}) + \cdots + (O^{\alpha'\alpha}\bar{\lambda}_{\alpha})(O^{\beta'\beta}\bar{\rho}_{\beta})\cdots(O^{\delta'\delta}\bar{\nu}_{\delta}) \quad (500)$$

If the O's are anti-unitary, so is O.

We next consider the application of these constructions to obtaining operators on $\mathcal{H}$. Recall that the non-interacting Hilbert space $\mathcal{H}$ is the tensor product of certain Fock spaces based on one-particle Hilbert spaces, e.g.,

$$\mathcal{H} = \mathcal{F}(H_D) \otimes \mathcal{F}(H_M) \otimes \mathcal{F}(H_{CKG}) \quad (501)$$

where $\mathcal{F}$ denotes the operation of taking the (symmetric or antisymmetric, as appropriate) Fock space.

Consider first the unitary or antiunitary operators U_P ($P \in \mathcal{P}$) which arise from the Poincaré group. These operators are defined originally on the one-particle Hilbert spaces. Their action is first extended to the many-particle Hilbert spaces via (499) or (500), and then to the Fock spaces via (497). Finally, these operators are defined on $\mathcal{H}$ via (499) or (500). Thus, we obtain a representation of the Poincaré group $\mathcal{P}$ on $\mathcal{H}$. For $P \in \mathcal{P}$, we write the corresponding operator on $\mathcal{H}$ as U_P. (No confusion will result from this duplicity of notation.) Note that all the operators U_P on $\mathcal{H}$ are norm-preserving, and that U_P is antilinear if and only if P reverses time orientation, linear otherwise. The energy, momentum, and angular momentum operators on $\mathcal{H}$ are obtained by considering the U_P's which differ infinitesimally from the identity (see Eqn. (222).) Similarly, the charge-reversal operator C is defined, first, on the one-particle Hilbert spaces, and then extended successively to the many-particle spaces, to the Fock spaces, and finally to $\mathcal{H}$. The resulting operator on $\mathcal{H}$ is again denoted by C.

Another operator of interest is the total charge operator, Q. On our real Hilbert spaces (which represent neutral particles), H_{RKG} and H_M, $Q = 0$. On the complex Hilbert spaces, H_{CKG} and H_D, Q takes one of the two forms

$$Q = eP^{-} - eP^{+} \quad (502)$$

$$Q = eP^+ - eP^- \tag{503}$$

where P^+ and P^- are the projection operators onto positive-frequency and negative-frequency parts, respectively. The choice between (502) or (503) is based essentially on convention. One uses (502) when the "particles" (positive-frequency solutions) have negative charge, and (503) when they have positive charge. (For example, one could perfectly well dictate that the positron is the "particle", and the electron the antiparticle. It is conventional, however, to make the assignments the other way. This choice his no physical consequences.) Note that, in every case, Q is Hermitian. We extend Q from the one-particle to the many-particle Hilbert spaces via (499), to the Fock spaces via (497), and to $\mathcal{H}$ via (499). The result is a Hermitian charge operator, Q, on $\mathcal{H}$.

Finally, we consider the creation, annihilation, and field operators. These operators, in distinction to the others defined above, are first defined on the Fock spaces rather on the one-particle Hilbert spaces. Suppose, for example, that $\mathcal{H}$ is given by (501), and we are considering the field operator $\underline{\phi}(f)$ on $\mathcal{F}(H_{CKG})$. Now consider the following triple of operators: $\mathbb{I}$ (the identity) on $\mathcal{F}(H_D)$, $\mathbb{I}$ on $\mathcal{F}(H_M)$, and $\underline{\phi}(f)$ on $\mathcal{F}(H_{CKG})$. This triple defines, via the construction (499), an operator on $\mathcal{H}$. In this way, the creation, annihilation, and field operators are extended from a single Fock space to $\mathcal{H}$. The resulting operators on $\mathcal{H}$ will, as usual, be denoted by the same symbols as the corresponding operators on the single Fock spaces.

All of the commutators and other relations between these various operators on $\mathcal{H}$ are simple to evaluate. We give a few examples. The operators associated with the Poincaré group satisfy, of course,

$$U_P U_{P'} = U_{PP'} \tag{504}$$

The U_P's leave invariant the charge of an element of $\mathcal{H}$ if P does not reverse the orientation, and reverse that sign if P does reverse time orientation. Hence, we have

$$U_P Q U_P^{-1} = \pm Q \tag{505}$$

with the plus sign if P does not reverse time orientation, the minus sign otherwise. Of course, charge-reversal reverses the sign of charge:

$$CQC^{-1} = -Q \tag{506}$$

The commutators between the field operators (and the creation and annihilation operators) within one Fock space are the same for $\mathcal{H}$ as for the original Fock space. That is, for example, we have

$$\left[\underline{\phi}(f), \underline{\phi}^*(g)\right] = \frac{\hbar}{2i} D(f,g)\mathbb{I} \tag{507}$$

on $\mathcal{H}$. Similarly, the adjoint relations between these operators are unchanged in the passage to $\mathcal{H}$. Field operators (as well as creation and annihilation operators) which act on different Fock spaces commute. For example,

$$\left[\underline{\phi}(f), \underline{A}(f^a)\right] = 0 \tag{508}$$

Finally, we consider the relationship between the field operators and C, Q, and U_P. Once again, everything is straightforward, so a single example will suffice. Consider H_{CKG}, so

$$\begin{aligned} \underline{\phi}(f) &= \hbar C(\sigma^-(f)) + \hbar A(\overline{\sigma^+(f)}) \\ \underline{\phi}^*(f) &= \hbar A(\overline{\sigma^-(f)}) + \hbar C(\sigma^+(f)) \end{aligned} \tag{509}$$

Let P be a restricted Poincaré transformation. Then, from (497), (499), (100), and (101),

$$U_P \underline{\phi}(f) U_P^{-1} = \hbar C(U_P \sigma^-(f) + \hbar A(\overline{U_P \sigma^+(f)}) \tag{510}$$

Similarly, if C is a particle-reversal operator, then

$$\begin{aligned} C\underline{\phi}(f)C^{-1} &= \hbar C(C\sigma^-(f)) + \hbar A(\overline{C\sigma^+(f)}) \\ &= \hbar C(\sigma^+(f)) + \hbar A(\sigma^-(f)) = \underline{\phi}^*(f) \end{aligned} \tag{511}$$

For the charge operator Q, we note that $\underline{\phi}(f)$ creates an antiparticle (say, with positive charge) and annihilates a particle. Hence, the total change in the charge effected by $\underline{\phi}(f)$ is just $-e$ times the norm of $\sigma^-(f)$ plus $-e$ times the norm of $\sigma^+(f)$. Thus,

$$[Q, \underline{\phi}(f)] = -e \left(\|\sigma^+(f)\| + \|\sigma^-(f)\| \right) \mathbb{I} \tag{512}$$

Clearly, the list of operators in this subject is almost infinite. Roughly speaking, any two operators in this list have a relationship which is simple, straightforward to derive, and easy to interpret physically.

34 Electromagnetic Interactions

In Sect. 29 (see Eqn. (455)) we wrote down an expression for the S-matrix in terms of an (unknown) operator field $\underline{K}(x)$ on Minkowski space. Of course, this formula gives practically no information about the scattering unless one knows $\underline{K}(x)$. One imagines that the actual $\underline{K}(x)$ which describes physical processes in the real world can be written as the sum a certain number of terms (e.g., the electron-photon interaction, the nucleon-photon interaction, the π-meson-photon interaction (electromagnetic interactions), the π-meson-nucleon interaction (strong interactions), the electron-neutrino interaction, the π-meson-neutrino interaction (weak interactions), etc.) There are at least some experimental situations in which one single term dominates all the others. One attempts to obtain an expression for this term using physical arguments and trial and error. That is to say, one makes a reasonable guess for the term in $\underline{K}(x)$, and compares the theoretical consequences of that guess (via (455)) with experiment. The hope is that one can, in this way, isolate and study each term, and then, by adding the well-established terms, obtain a reasonable approximation to the "actual" $\underline{K}(x)$ which is operative in Nature. We shall here merely illustrate the general idea by writing down and discussing a few of the $\underline{K}(x)$'s associated with the interaction of charged particles with the electromagnetic field.

We begin with the simplest case: the interaction of a complex Klein-Gordon field with the Maxwell field, e.g., the interaction of $\pi^{\pm}$-mesons with photons. In this case, we would take for our non-interacting Hilbert space

$$\mathcal{H} = \mathcal{F}(H_{CKG}) \otimes \mathcal{F}(H_M) \tag{513}$$

What should we choose for the operator field $\underline{K}(x)$ on $\mathcal{H}$? In this case, we have an important physical clue: we know what the classical "interaction energy density" is between a classical Klein-Gordon field and a classical Maxwell field, namely

$$K(x) = \frac{ie}{2\hbar}\left(\phi(x)\nabla_a\bar{\phi}(x) - \bar{\phi}(x)\nabla_a\phi(x)\right)A^a(x) \tag{514}$$

(Note that it is only an integral of (514) which has meaning, for we have the freedom to add a gradient to the vector potential. Appropriate integrals are gauge-invariant, however, because the first integral is, as we have seen, divergence-free.) In (514), e is a constant. Using the discussion of Sect. 30, and the fact that $\underline{K}(x)$ has dimensions of energy density (mt^{-3}), we see that the coupling constant e has dimensions $m^{1/2}t^{1/2}$, whence $e^2/\hbar$ is dimensionless. In order to obtain eventual agreement with experiment, it will, of course, be necessary to set this constant to $1/137$.

While (514) is a perfectly nice scalar field (on Minkowski space) constructed out of a complex Klein-Gordon field and a Maxwell field, it is, unfortunately, just that — a scalar field rather than an operator field. The expression (514) is just not the right sort of mathematical object to be $\underline{K}(x)$. Now comes the "transition from classical to quantum theory." Roughly speaking, what we propose to do is to replace the classical fields in $K(x)$ (Eqn. (514)) by the corresponding field operators to obtain $\underline{K}(x)$. Unfortunately, this replacement is not so simple and unambiguous as it may appear at first sight.

By what operator should we replace $\phi(x)$? Our Klein-Gordon field operator, $\underline{\phi}(f)$, depends on test fields in Minkowski space, and not on points of Minkowski space. What one would like to do is define an operator field $\underline{\phi}(x)$ by

$$\underline{\phi}(x) = \lim_{f \to \delta_x} \underline{phi}(f) \tag{515}$$

where δ_x denotes a δ-function located at the point x. But will the limit in (515) exist? The answer, as we have seen earlier, is no. We could still regard $\underline{\phi}(x)$ as an operator-valued distribution (i.e., a linear mapping from test functions to operators on $\mathcal{H}$ — that, after all, is what $\underline{\phi}(f)$ is), but such an attitude again leads to difficulties. Eqn. (514) will require us to take products of such operator-valued distributions, but the ability to take products is precisely what is lost in the transition from functions to distributions. That is to say, products of distributions are not in general well-defined. (This is a genuine and serious problem — not to be confused, for example with the standard complaints about use of the Dirac δ-function.) In short, we are stuck. There is no meaning which can be given to (515) which would be appropriate for replacement in (514).

We adopt the following attitude. We leave the problem of the nonexistence of limits such as (515) unresolved for the time being. We permit ourselves to

manipulate such quantities formally, as though the question of the limits had never arisen. This marks the third (and, mercifully, the last) of the mathematical problems associated with this formalism. For emphasis, we list these problems again:

1. The question of the convergence of the infinite sum of operators in (455).
2. The question of the convergence of the integrals (over all of Minkowski space) of operators in (455).
3. The nonexistence of the "δ-function limits" of field operators used in obtaining $\underline{K}(x)$.

The situation will look less gloomy shortly. (I find it hard to believe that the "ultimate", mathematically acceptable, quantum field theory will result from a brute-force attack on these problems.)

We now have $\underline{\phi}(x)$ to replace $\phi(x)$. Naturally, we replace the classical complex-conjugate field, $\bar{\phi}(x)$, by the adjoint operator, $\underline{\phi}^*(x)$. What about the derivative terms? Let p^a be a vector at the point x, and extend p^a to a constant vector field on Minkowski space. We then define

$$p^a \nabla_a \underline{\phi}(x) = - \lim_{f \to \delta_x} \underline{\phi}(p^a \nabla_a f) \tag{516}$$

(See Eqn. (145).) Similar remarks concerning non-existence of limits apply.

We must next select an operator to replace the vector potential, $A_a(x)$, in (514). Ideally, one would like to define $ulA^a(x)$ by

$$p^a \underline{A}_a(x) = \lim_{f \to \delta_x} \underline{A}(p^a f) \tag{517}$$

where p^a is a constant vector field, and $\underline{A}(\ )$ is the field operator (Sect. 14) for the vector potential. Unfortunately, this won't work, for $\underline{A}(f^a)$ is only defined for test fields f^a which can be written as the sum of a divergence-free field and a gradient; $p^a f$ cannot be written in this form in general. The simplest way of overcoming this difficulty is as follows. First note that the commutator of the vector potential operators ((385) and (386)) is well-defined whether or not the test fields, f^a and g^a, can be written as the sum of a divergence-free field and a gradient. In fact, it is only these commutators which will enter the S-matrix. Hence, we can work with vector potential operators, $\underline{A}_a(x)$, and use for the commutators (385).

We now have an operator equivalent for each term in (514). We must now face the next problem: in what order should the operators be placed? This difficulty does not arise, of course, in the classical theory, because the classical fields may be placed in any order. We consider the most general linear combination:

$$\underline{K}(x) = \frac{ie}{2\hbar} \left(a \underline{\phi} \nabla_a \underline{\phi}^* + b(\nabla_a \underline{\phi}^*) \underline{\phi} + c \underline{\phi}^* \nabla_a \underline{\phi} + d(\nabla_a \underline{\phi}) \underline{\phi}^* \right) \underline{A}^a \tag{518}$$

where a, b, c, and d are real numbers. Taking the Hermitian conjugate of (518),

$$\underline{K}(x) = -\frac{ie}{2\hbar} \left(a(\nabla_a \underline{\phi}) \underline{\phi}^* + b \underline{\phi}^* \nabla_a \underline{\phi} + c(\nabla_a \underline{\phi}^*) \underline{\phi} + d \underline{\phi} (\nabla_a \underline{\phi}^*) \right) \underline{A}^a \tag{519}$$

We see that the Hermiticity of (518) requires

$$a = -d \qquad b = -c \tag{520}$$

Further information about the coefficients is obtained from the experimental fact that electromagnetic interactions are invariant under charge reversal. From (518):

$$C\underline{K}(x)C^{-1} = \frac{ie}{2\hbar}\left(a\underline{\phi}^*\nabla_a\underline{\phi} + b(\nabla_a\underline{\phi})\underline{\phi}^* + c\underline{\phi}\nabla_a\underline{\phi}^* + d(\nabla_a\underline{\phi}^*)\underline{\phi}\right)\underline{A}^a \tag{521}$$

Thus, invariance under charge reversal requires one of the following two alternatives:

$$a = c \qquad b = d \qquad C\underline{A}^aC^{-1} = \underline{A}^a \tag{522}$$

$$a = -c \qquad b = -d \qquad C\underline{A}^aC^{-1} = -\underline{A}^a \tag{523}$$

We choose (523) for two reasons: (i) it is more reasonable on physical grounds to have the vector potential reverse sign under charge reversal (for classical electromagnetic fields reverse sign when the signs of all charges are reversed), and (ii) with this choice, $\underline{K}(x)$ reduces, in the classical limit, to the classical expression (514). Thus, we arrive at the interaction:

$$\underline{K}(x) = \frac{ie}{4\hbar}\left(\underline{\phi}\nabla_a\underline{\phi}^* + (\nabla_a\underline{\phi}^*)\underline{\phi} - \underline{\phi}^*\nabla_a\underline{\phi} - (\nabla_a\underline{\phi})\underline{\phi}^*\right)\underline{A}^a \tag{524}$$

Eqn. (524) describes an interaction which is invariant under charge reversal. Note, furthermore, that if P is any Poincaré transformation, then

$$U_P\underline{K}(x)U_P^{-1} = \underline{K}(Px) \tag{525}$$

Since $\underline{K}(x)$ is integrated over all of Minkowski space (Eqn. (455)), the final S-matrix will commute with each U_P. Thus, our interaction conserves the quantities associated with the infinitesimal generators of energy, momentum, and angular momentum. The interaction is also invariant under parity and time reversal. Note, furthermore, that we have

$$[\underline{K}(x), \underline{K}(y)] = 0 \tag{526}$$

for $x - y$ spacelike, for when $x - y$ is spacelike, any two operators in (524) commute with each other. Finally, the total charge operator, Q, commutes with $\underline{K}(x)$, for $\underline{A}_a(x)$ commutes with Q, $\underline{\phi}(x)$ decreases the total charge by $2e$, while $\underline{\phi}^*(x)$ increases it by $2e$.

To summarize, (524) is an interaction which is invariant under parity, time, and charge reversal, and conserves charge, energy, momentum, and angular momentum.

As a second example, we discuss the interaction of photons with electrons and positrons. As before, we begin with the classic interaction energy density:

$$K(x) = e\left(\xi^A\bar{\xi}^{A'} + \eta^{A'}\bar{\eta}^A\right)A_{AA'} \tag{527}$$

The classical Dirac fields are to be replaced by the following operators:

$$\begin{aligned} p_A \underline{\xi}^A(x) &= \lim_{f \to \delta_x} \frac{1}{2} \left[\underline{\psi}(f p^A, f \bar{p}_{A'}) - i \underline{\psi}(i f p^A, -i f \bar{p}_{A'}) \right] \\ \bar{p}^{A'} \underline{\eta}_{A'}(x) &= \lim_{f \to \delta_x} \frac{1}{2} \left[\underline{\psi}(f p^A, f \bar{p}_{A'}) + i \underline{\psi}(i f p^A, -i f \bar{p}_{A'}) \right] \end{aligned} \tag{528}$$

where p^A is a constant spinor field. Note that e in (527) again has dimensions $m^{1/2}t^{1/2}$. The classical complex-conjugate fields, $\bar{\xi}^{A'}$ and $\bar{\eta}_A$, are to be replaced by the Hermitian conjugates, $\underline{\xi}^{*A'}$ and $\underline{\eta}^*_A$, respectively. In this case, the problem of factor ordering is not resolved by the requirement that $\underline{K}(x)$ be Hermitian: this condition is satisfied for any factor ordering. However, this electromagnetic interaction should be invariant under charge reversal. We have

$$\begin{aligned} C \underline{\xi}^A C^{-1} &= \underline{\eta}^{*A} \\ C \underline{\eta}_{A'} C^{-1} &= \underline{\xi}^*_{A'} \end{aligned} \tag{529}$$

What should we adopt for the behavior of the vector potential operator, $\underline{A}_a(x)$, under charge reversal? We have already decided, for the meson-photon interaction, to use

$$C \underline{A}_a C^{-1} = -\underline{A}_a \tag{530}$$

It is an important point that we must choose the same behavior for the present interaction. The reason is that, for the "actual" interaction which Nature obeys, $\underline{K}(x)$ will be the sum of the various interactions. If we use a different charge-reversal operator for each term which appears in this sum, then we will have no operator which commutes with the total $\underline{K}(x)$. In other words, the behavior of each type of particle under the various reversals must be fixed once and for all. One has, of course, freedom to choose that behavior, and this choice is based on obtaining operators which commute with as many terms in the "final" $\underline{K}(x)$ possible. Thus, using (529) and (530), we are led to adopt the expression

$$\underline{K}(x) = \frac{1}{2} e \left(\underline{\xi}^A \underline{\xi}^{*A'} - \underline{\xi}^{*A'} \underline{\xi}^A + \underline{\eta}^{A'} \underline{\eta}^{*A} - \underline{\eta}^{*A} \underline{\eta}^{A'} \right) \underline{A}_{AA'} \tag{531}$$

for the interaction.

Note that (530) is Hermitian, and that the resulting S-matrix commutes with the unitary operators which arise from the Poincaré group. Thus, (530) conserves energy, momentum, and angular momentum. By the same argument as before, $\underline{K}(x)$ commutes with the total charge operator Q. Finally, we note that, if $x - y$ is spacelike,

$$[\underline{K}(x), \underline{K}(y)] = 0 \tag{532}$$

This arises from the following facts: when $x - y$ is spacelike, any two boson operators commute, while any two fermion operators anticommute. But $\underline{K}(x)$ contains an even number of fermion operators. Since reversing the order of two boson operators gives a plus sign, and reversing the order of two fermion

operators gives a minus sign, the total number of minus signs will be even, and so we have (532).

Clearly a vast number of conceivable interactions could be written down using the pattern illustrated above. One first writes down a real scalar field constructed from the classical fields. One then replaces the classical fields by the corresponding operators. The factors must be ordered so that the resulting operator is Hermitian, and satisfies (532). Beyond that, the choice of factor ordering must be based on physical or aesthetic considerations, experiment, etc. We have merely discussed two possible interactions here in order to illustrate the method. (In fact, these are the two simplest, for one can rely heavily on classical theory as a guide.)

35 Transition Amplitudes

Suppose now that we have selected a particular $\underline{K}(x)$, and wish to work out its experimental consequences, using (455). The straightforward procedure — substituting $\underline{K}(x)$ into (455), and attempting to carry out the integrals and sum — turns out to be too difficult to carry out in practice. Instead, one adopts a more indirect approach — which leads ultimately to the Feynman rules. We shall not attempt to derive the Feynman rules, or even discuss the large volume of technical apparatus which has been developed to deal with (455). Instead, we merely indicate the general idea of the method.

Suppose first that we were able, in some way, to obtain the value of the complex number

$$(\tau, S\sigma) \tag{533}$$

for any two states $\sigma, \tau \in \mathcal{H}$. (The expression (533) is called the *transition amplitude* between the state σ and the state τ.) This information is, of course, completely equivalent to a knowledge of the S-matrix. In fact, it suffices to know (533) only for σ's and τ's drawn from a certain subspace of $\mathcal{H}$, provided this subspace is dense in $\mathcal{H}$. Let σ_0 denote the vacuum state in $\mathcal{H}$ and $C_1, C_2, \ldots, C_n$ any finite sequence of creation operators on $\mathcal{H}$. (One C_i might create a photon, another an electron, others mesons, etc.) We consider the state (element of $\mathcal{H}$)

$$C_1 C_2 \cdots C_n \sigma_0 \tag{534}$$

Clearly, the collection of all finite linear combinations of states of the form (534) is dense in $\mathcal{H}$. Hence, it suffices to evaluate

$$\begin{aligned}(C_1 \cdots C_n\sigma_0, SC'_1 \cdots C'_m\sigma_0) &= (C_1 \cdots C_n\sigma_0, C'_1 \cdots C'_m\sigma_0) \\ &\quad + \left(-\frac{i}{\hbar}\right) \int \mathrm{d}V_1 (C_1 \cdots C_n\sigma_0, \underline{K}(x_1)C'_1 \cdots C'_m\sigma_0) \\ &\quad + \frac{1}{2!}\left(-\frac{i}{\hbar}\right)^2 \int \mathrm{d}V_1 \int \mathrm{d}V_2 (C_1 \cdots C_n\sigma_0, T\left[\underline{K}(x_1), \underline{K}(x_2)\right] C'_1 \cdots C'_m\sigma_0) \\ &\quad + \cdots \end{aligned} \tag{535}$$

for any $C_1, \ldots, C_n, C'_1, \ldots, C'_m$. One now attempts to evaluate the various terms in the sum (535) individually. The first term is celled the 0^{th}-order interaction. It clearly vanishes unless $n = m$, and the C's and C''s create the same number of photons, the same number of electrons, etc. The second term is called the first-order interaction, etc. These orders correspond to the various orders of the Feynman diagrams.

The next step is based on the following observation. The operator $\underline{K}(x)$ is expressed in terms of the field operators, and can, therefore, be written in terms of creation and annihilation operators. Thus, each inner product in (535) is equal to the vacuum expectation value of some product of creation and annihilation operators. But we have seen (p. 89) that such expectation values can be reduced to products of the propagators. (One pushes the annihilation operators to the the right, using the commutators, until they reach σ_0, and there give zero.) Thus, each term in (535) can be written as an integral of a product of operators. The Feynman rules represent a technique for writing out such integrals directly without going through the algebra (p. 89). Since $\underline{K}(x)$ was expressed in terms of the "unsmeared" field operators, the propagators will be distributions. Thus, the evaluation of the S-matrix reduces to integrating certain products of distributions over Minkowski space. Of course, the integrals diverge. These are the divergences.

We illustrate these remarks with one example. Consider the second-order interaction with $\underline{K}(x)$ given by (531). Now, $\underline{K}(x)$ contains two lepton operators and one photon operator. Thus, the transition probability will vanish (in the second order) unless the number of outgoing leptons differs by no more than four from the number of ingoing leptons, and the number of outgoing photons differs by no more than two from the number of ingoing photons. Clearly, the higher the order of the interaction, the more possibilities for creating and annihilating particles. Because of the smallness of the coupling constant in this electromagnetic interaction, $e^2/\hbar = 1/137$, many experiments are adequately described by the first few orders.

> "Would it not be better to get *something* done, even though one might not quite understand *what*?"
>
> – J. L. Synge

Perspectives in Computation

Robert Geroch

March 15, 2016

Contents

1 Introduction

This book comprises the lecture notes for a course I taught for a small group of graduate students in physics at the University of Chicago in Winter quarter, 2005 and 2007. Our subject was the theory of "computation", broadly defined.

This is a fascinating field. It is full of open questions, many of which look deceptively easy but turn out to be extremely difficult. It is made all the richer by its position on the borderline between mathematics and physics. This is also an extremely active field, with an enormous literature. We shall take a brief walking tour through this subject, trying to pin down a few of the issues that struck me as the most illuminating and interesting. Our emphasis is on what might be called "conceptual issues" — the formulation of precise definitions, how questions are to be framed, what is possible or not possible in principle, etc — as opposed to other, more practical, issues. This — the conceptual background of computation — constitutes a coherent subject in its own right; and we focus on it in order to arrive, quickly and clear-headedly, at some of the central ideas of this field.

This volume should be understandable to readers who know what mathematics is (i.e., what a definition is; what a theorem is; etc); and, for the last half of the volume, who have a rudimentary knowledge of quantum mechanics. We require virtually no technical knowledge in either mathematics or

physics. Thus, the volume should be accessible to many graduate students (and even advanced undergraduates) in mathematics or physics.

The topics that we discuss may be divided into three broad classes: i) the idea of computability; ii) the notion of efficiency in the carrying out a computation; and iii) the possibilities for using quantum mechanics in the computation process. Each of these is an entire field all by itself; and each would require for a proper treatment several courses and several volumes. It is thus well beyond our scope to provide an in-depth survey of even the smallest portion of these subjects.

The first collection of topics — involving the idea of computability — is covered in Sects. 2-8. Sects. 2 and 3 introduce some preliminary notions: the idea of strings over a character set (the currency of the computation process); and the idea of a problem (a "sequence of questions", expressed in terms of such strings). Sects. 4 and 5 introduce the notion of computability — of having an algorithm (expressed as a suitable computer program) to answer each of the questions in succession. It turns out that there exist problems that are not computable. That is, each question in the sequence has a definite answer, but there is no single algorithm for answering them all. In fact, one can write down explicit examples of such problems. (The strategy is to construct the problem by using the computing language itself.) These topics are discussed in Sects. 6 and 7.

The second collection of topics — involving computational efficiency — is covered in Sects. 10-13. Fix a computable problem, together with an algorithm to compute that problem. The key idea (Sect 9) is to introduce a number representing the "difficulty" of that computation, i.e., roughly, the number of steps required to carry it out. The difficulty of a problem as a whole is thus a function of the input string. Sect. 10 contains two remarkable results, due to Blum, which serve to illustrate how computational difficulty works. The first result asserts that there exist essentially "arbitrarily difficult" problems (suitably defined). The second is to the effect that there exist problems for which there is no "best" algorithm: Given any program P that computes such a problem, there exists another program that computes the same problem, but more efficiently than P does. This last result seems to destroy any hope of ever assigning, to every problem, an intrinsic difficulty.

This whole idea of computational difficulty suffers from a, potentially serious, drawback. Whereas the notion of computability is essentially language-independent, that of difficulty is not — there are, for example, intrinsically

inefficient languages. Better would be a notion of difficulty that attaches to the problem and the method for its solution (but not directly to the language in which that method is written). Sects 11 and 12 represent a proposal to capture such a notion. The strategy is to look for a language that is "as simple as possible, without sacrificing efficiency". It might be interesting to look for a theorem to the effect that this proposed language really does what it is designed to do. Finally, Sect. 13 introduces the idea of incorporating probability into the computation process. It turns out that there is a natural notion of a probabilistic program; of such a program's computing a problem; and of the difficulty function of such a program. Probability does not add anything to computability: The probabilistically computable problems turn out to be identical to those computable by a regular program. But — remarkably enough — it remains an open question as to whether the use of probability can effectively reduce the difficulty of a computation.

The third collection of topics — involving the use of quantum mechanics in the computation process — is covered in Sects. 14-21. This is a very active area of research. Rather than trying to summarize this large body of work, we will focus on two — rather narrowly drawn — issues. The first is to translate, into mathematics, the physics of utilizing quantum mechanics in the computational process. The second is to examine the issue of whether quantum mechanics can enhance computational efficiency. Sect. 14 is a very brief review of quantum mechanics — for people who already know quantum mechanics. Sect. 15 is a self-contained, four-page exposition of one example, called the Grover construction, in which quantum mechanics seems to hold out the prospect of enhanced efficiency. Those interested only in the role of quantum mechanics in computation may wish to start with this section. Sect. 16 discusses various issues — some of which turn out to be rather subtle — that arise in connection with the Grover construction. These three sections, 14-16, are intended as an informal introduction to the idea of using quantum mechanics as an aid to computation. They serve as motivation for Sects. 17-19, the key sections of this volume. In Sect. 17, we specify precisely what we mean by a "quantum-assisted computation": We introduce a certain precise (mathematical) computer language (a language that involves certain Hilbert spaces and operators on those spaces). This language is designed to reflect what could be done, in the laboratory, using quantum systems. In Sect. 18, we show that, provided matters are set up correctly, quantum-assistance, as here defined, will not enable us to "solve" any otherwise-noncomputable

problems. This comes as no surprise (although the "setting up" must be done with some care). In Sect. 19, we define the difficulty function for a quantum-assisted computation.

We now have, for any given problem, the notion of its being computed, and the difficulty in doing so, by a regular program; and also by a quantum-assisted program. We have a "level playing field" — a well-posed mathematical framework within which we may compare quantum-assistance with the lack thereof. There are indications that, for certain problems, quantum mechanics may be utilized in the course of the computation in such a way to enhance the efficiency of that computation. And yet, these indications notwithstanding:

> We have today not one single example of a problem, together with a quantum-assisted computation of that problem, such that we can actually *prove* that at least the same efficiency cannot be achieved already without using quantum mechanics.

This is, to my mind, a remarkable state of affairs. The basic obstacle here is, not any uncertainty about how quantum mechanics works in the computation process, but rather the lack of good lower bounds on the difficulty of *regular* computations. Such bounds seem to be lacking for even the simplest problems, e.g., that of multiplying two integers! These issues are discussed in Sects. 20 and 21. In Sect. 20, we prove a theorem to the effect that, if there does exist a problem for which quantum mechanics produces any gain in efficiency, then that gain can never be more than exponential. Finally, Sect. 21 discusses a few ideas about how one might prove (or disprove) that quantum mechanics has the ability to enhance computational efficiency.

I hope that these lecture notes will provide a brief glimpse into this most fascinating field.

2 Characters and Strings

Fix a finite set, $\mathcal{C}$, having at least two elements. This $\mathcal{C}$ will be called the *character set*; and its elements the *characters*. We shall normally introduce various symbols to denote the various elements of $\mathcal{C}$. For example, $\mathcal{C}$ might have just two elements, and these might be denoted $0, 1$. Or, $\mathcal{C}$ might consist of twenty-six elements, with these denoted $a, b, \cdots, z$; or 36 elements, denoted

$a, b, \cdots, z, 0, 1, \cdots, 9$. Or, as a final example, $\mathcal{C}$ might consist of 256 elements, denoted by the various ASCII characters.

The underlying choice of character set makes no significant (i.e., no interesting) difference to anything that follows; although a poor choice can turn out to be inconvenient. Eventually (but not right now) we shall allow ourselves to be a little sloppy as to exactly what our character set is.

Fix a character set, $\mathcal{C}$. By a *string* over $\mathcal{C}$, we mean a finite, ordered list of characters. Examples of strings, for the character sets above, are: "001010", "unblowupable", "$3dafrq$", and "$\$ = log8+\} - -r\&C$", respectively. The empty list of characters, which we denote $\emptyset$, is also allowed as a string; and it is called the *empty string*. [Thus, in order to avoid confusion, we shall avoid denoting any character by "$\emptyset$".]

The set of strings, over a given character set $\mathcal{C}$, will be denoted $\mathcal{S}$ (or $\mathcal{S}_{\mathcal{C}}$, if there is a chance of confusion as to what the character set is). Note that $\mathcal{S}$ is an infinite set; in contrast to $\mathcal{C}$, which is a finite set. A critical idea in this subject is to pass to "the infinite" in a careful, controlled way. Here, that passage is taking place in the construction of $\mathcal{S}$ from $\mathcal{C}$. This is something that we carry out directly, as opposed to letting $\mathcal{C}$ already be infinite on its own, in any old manner that it chooses.

Part of the reason why the "choice of character set makes no significant difference" is that it is possible to pass from one character set to another. For example, let $\mathcal{C} = \{0, 1\}$ and $\mathcal{C}'$ the ASCII character set. Then "writing a byte as eight bits" provides a mapping from $\mathcal{S}_{\mathcal{C}'}$ to $\mathcal{S}_{\mathcal{C}}$. For example, string "$ab$" $\in \mathcal{S}_{\mathcal{C}'}$ might be sent to the string "0000100100001010" $\in \mathcal{S}_{\mathcal{C}}$. Unfortunately, this mapping is not invertible: It is not true that every string over $\mathcal{C}$ arises in this way from *some* string over $\mathcal{C}'$. [Indeed, a string over $\mathcal{C}$ *does* arise in this way if and only if the number of characters of which it consists is divisible by eight.]

Next, fix a character set, $\mathcal{C}$; and fix also an ordering of the characters in that set (i.e., a choice of a "first" character, a "second" character, etc. through all the characters in the set). For example, $\mathcal{C}$ might be the lowercase Latin letters (the set of 26 characters, $\{a, b, \cdots, z\}$), and the ordering might be alphabetical. Having made these choices, we now construct an ordering also of the set $\mathcal{S}$ of strings over $\mathcal{C}$, in the following manner. First is the empty string, $\emptyset$; then all the one-character strings, in the ordering of the characters; then all two-character strings, in dictionary ordering; then all three-character strings, etc. So, for instance, in the example above the ordering

of $\mathcal{S}$ is: $\emptyset$, "a", $\cdots$, "z", "aa", "ab", $\cdots$, "az", "ba", "bb", $\cdots$, "zz", "aaa", $\cdots$. Now assign, to these strings so ordered, successive integers, beginning with the integer 1. In this way, we set up a correspondence between the set $\mathcal{S}$ and the set $\mathcal{Z}^+$ of positive integers. In short, strings are really just positive integers, in thin disguise. Indeed, we shall allow ourselves to speak of "the n^{th} string", by which we shall mean the n^{th} string in this ordering, where some fixed ordering of the character set $\mathcal{C}$ is implicit (or explicit). When dealing with mathematical issues (such as the manipulation recursive functions), it is sometimes more convenient to stick entirely with the integers, ignoring character sets and strings altogether. But the strings seem better adapted to dealing with physical issues.

The reason that we required that the character set contain at least two elements is that, for $\mathcal{C}$ having but a single element, the length of the n^{th} string grows linearly with n (rather than logarithmically, when $\mathcal{C}$ contains two or more characters). This behavior would be inconvenient when we come to discuss computational difficulty.

The ordering above leads to a different way to pass from strings over one character set to those over another. First order both character sets (in any way whatever); and then identify, for $n = 1, 2, \cdots$, the n^{th} string over the first character set with the n^{th} string over the second. Consider, for example, the character sets given by $\{0, 1\}$ and $\{a, b, \cdots, z\}$, in the indicated orderings. Then the string "1101" over the first character set would be identified, in this manner, with the string "ab" over the second character set. Note that (in contrast with the earlier method) this really is a correspondence between the two sets of strings, i.e., this mapping from $\mathcal{S}_{\mathcal{C}}$ to $\mathcal{S}_{\mathcal{C}'}$ is one-to-one and onto.

3 Problems

Fix a character set, $\mathcal{C}$. A *problem* is a mapping $\mathcal{S} \stackrel{\pi}{\to} \mathcal{S}$, i.e., a mapping from strings over $\mathcal{C}$ to strings over $\mathcal{C}$. Here is an example of a problem:

> **Example.** Let the character set have thirteen elements, $0, 1, \cdots, 9$, y, n, f, and let the mapping π be the following. If the string $S \in \mathcal{S}$ is an integer greater than or equal to two (i.e., if it is not $\emptyset$ or "1", does not contain the characters "y", "n", or "f", and does

not have an initial character "0")[1] then let $\pi(S)$ be "y" ["yes"] if the integer S is prime, and "n" ["no"] if that integer is not prime. If the string S is not an integer greater than or equal to two, then set $\pi(S) =$ "f" ["forget it"]. This is indeed a mapping $\mathcal{S} \stackrel{\pi}{\to} \mathcal{S}$, and so is a problem.

Note that, in the example above, we are actually only interested in certain strings (namely, those that represent integers greater than or equal to two). But we cannot confine ourselves simply to these strings, for, by definition of a "problem", the mapping must apply to *all* strings. So we send the ones we aren't interested in to the trash ["f"]. It is convenient to set things up in this way. Suppose, for example, that we had defined a problem to be a mapping from a mere subset of $\mathcal{S}$ to $\mathcal{S}$. Then, e.g., it would be false that the composition of two problems is a problem. Even worse, we would have to confront eventually the issue of how we shall determine whether or not a given string is in the "certain subset". The present definition — requiring that the mapping π have domain all of $\mathcal{S}$ — puts the burden of sorting all this out back on the mapping π itself (where, as we shall see, it belongs).

In the example above, the problem is really the question "Is integer S prime, or is it not?" Note that we could ask essentially this same question by any number of other maps (i.e., by any number of other problems). For instance, we could eliminate "y", "n", and "f" from the character set, and then encode the answer as a digit (e.g., "0" for "prime"; "1" for "integer ≥ 2 but not prime"; and "2" for "not an integer ≥ 2"). We could also modify the input. For example, we could let the character set consist of the twenty-six lower-case Latin letters, impose alphabetical ordering, and let π ask whether, given string S (say, the n^{th} string in the induced ordering on $\mathcal{S}$) the integer n is prime. Thus, we see that a given question may give rise to many problems.

We emphasize that a problem is a *map*, and not the process by which we arrived at that map. Consider, for example, the following problem:

Example Let the character set again be $0, 1, \cdots, 9, y, n, f$, and let $\pi(S)$ be "f" if S does not represent an integer greater than or

[1] In order to avoid having to say all this repeatedly, let us agree, here and hereafter, on the following definition: Over any character set that includes the digits, $0, 1, \cdots, 9$, a string will be called an *integer* if it contains only those ten characters, is not $\emptyset$, and (unless it is actually the string "0") does not have initial character "0".

equal to two, "y" if it does represent such an integer, and that integer is the largest prime, and "n" if it does represent such an integer, and that integer is not the largest prime.

This is the same *problem* (i.e., the same map) as that which sends string S to "n" if S represents an integer greater than or equal to two, and "f" otherwise. [This follows, since there is no largest prime.] But this is a very different characterization of the problem than our earlier one. Thus, we see that several apparently different questions may give rise to precisely the same problem.

Think of a problem π as a "broad question"; of a particular value for the argument S as a "specific instance" of that broad question; and of $\pi(S)$ as the answer to that question in that instance. Thus, a problem represents the answers to an infinite number of questions (since there is an infinite number of possible input strings). The prime problem above illustrates this point. But it is also possible to design a problem that answers a single question. For example, consider

> **Goldbach's Conjecture:** Every even integer $n \geq 4$ is equal to a sum of two primes.

Let the character set be $0, 1, \cdots, 9, y, n, f$, as above, and let, for each string $S \in \mathcal{S}$, $\pi(S) =$ "y" if Goldbach's conjecture is true, and $\pi(S) =$ "n" if Goldbach's conjecture is false. [Note that π doesn't care what S is. No law says it must.] Since the conjecture above is, presumably, either true or false, this is indeed a problem. However, this problem is *either* the problem $\pi'(S) =$ "y" for all S; *or* the problem $\pi''(S) =$ "n" for all S. But these are both rather uninteresting problems. Thus, even though the original question ("Is Goldbach's conjecture true?") is interesting, translating it in this way into a problem yields what is guaranteed to be a pretty boring problem. Of course, to know whether π is π' or π'' would be interesting.

Here are some other candidates for problems.

> **1.** Let the character set be $\{0, 1, \cdots, 9, f\}$, and let $\pi(S)$ be "f" if S is not a positive integer, and the S^{th} digit in the decimal expansion of the number π if S is a positive integer. This is a problem.

2. Let the character set be the same, but again let $\pi(S)$ be "f" if S is not a positive integer. But if S is a positive integer, let $\pi(S)$ be a string of the form {digits of a positive integer i}f{digits of a positive integer i'}, where i/i' is within 10^{-S} of the number π. This is not a problem (since the actual mapping has not been specified). Rather, this is a description of a class of problems. There does indeed exist a problem in this class (e.g.: expand the number π decimally, as in the first example; stop as soon as you reach a rational within 10^{-S} of the number π; and finally reduce that fraction to lowest terms).

3. Let the character set be $\{0, 1, \cdots, 9, y, n, f\}$. Let $\pi(S)$ be "f" if S is not an integer ≥ 4. If S is an integer ≥ 4, let $\pi(S)$ be "y" if S is a counterexample to the Goldbach conjecture, and "n" if it is not. This is a problem. The question of whether Goldbach's conjecture is true is the question of whether or not the problem π is equal to a suitable simple problem (which always answers "f" or "n").

4. Let the character sets consist of the upper-case and lower-case Latin letters, together with an appropriate set of punctuation marks (period, comma, semicolon, question mark, exclamation mark, etc). Let $\pi(S)$ be "yes!" if S is the Gettysburg address, and "no" otherwise. This is a problem.

5. Let the character set be any ordered set that includes at least the ten digits (not necessarily in their natural order). Let π send string S over this character set to that integer n which is such that S is the n^{th} string. This is a problem.

It is interesting to note that there has taken place a progression to ever greater levels of the infinite. The character set is finite; the set of strings over that character set is countably infinite; and, finally, the set of problems on those strings is uncountably infinite. We pause to give a proof of this last assertion because it illustrates a method, called a diagonal argument, that we shall use several times later. Fix $\mathcal{C}$, and suppose, for contradiction, that

we had a countable collection of problems, $\pi_1, \pi_2, \pi_3, \cdots$, that exhausted all problems on this character set. We now introduce a new problem, π, as follows. Say, one of the characters is "a". On the n^{th} string, S_n, set $\pi(S_n) = \emptyset$ if $\pi_n(S_n) =$ "a"; and $\pi(S_n) =$ "a" otherwise. Then this problem π, so constructed, is not equal to π_n for any n, since by construction $\pi(S_n) \neq \pi_n(S_n)$. Thus, the list $\pi_1, \pi_2, \cdots$ could not have been exhaustive — a contradiction.

Finally, we remark that this notion of a problem is rather robust. For virtually the entirety of the remaining discussion, we shall have before us some problem, as here defined, or other.

4 Computability

Fix a character set, $\mathcal{C}$. We next wish to introduce the notion of computability of a problem over this character set.

Roughly speaking, a problem $\mathcal{S} \xrightarrow{\pi} \mathcal{S}$ is computable provided there exists a computer that, when run with any given input string S, will ultimately halt, displaying at that point as output precisely the string $\pi(S)$. But what is a "computer"? We cannot take this to mean a physical computer, because no such computer is ever capable of solving any problem. My desktop, for example, has a hard drive with capacity of only 10 GB. Thus, if I let the character set be, say, ASCII, and let the input string S consist, say, of 10^{11} characters, then surely this computer will be unable run with this input.

More promising would be to consider, rather than a physical computer, a computer language. Consider Fortran[2]. A given Fortran program has no space limitations whatever associated with it. You begin by purchasing some physical computer, and running the given program on it. Then if, during the course of the calculation, it turns out that there is insufficient space to complete that calculation, you will be invited to purchase a larger computer and rerun the program on it. So, let us call a problem $\mathcal{S} \xrightarrow{\pi} \mathcal{S}$ "Fortran-computable" if there exists a Fortran program with a single, initial, "Input" instruction (allowing the user to input some string, S), a single, final, "Print" instruction (allowing the program to display to the user some

[2]We shall not be concerned here with details of any specific computer languages. In particular, we take "Fortran" as a generic term, which describes languages having such commands as "Set $x = ...$", "If $(\cdots)$, go to ...", "Do, for $I = 1, n\{\cdots\}$ Next", "Print ...", "Cat x, y", etc.

final string), having the following property: For every choice of the input S, the program ultimately halts (as opposed, e.g., to getting caught in an infinite loop), having printed precisely the string $\pi(S)$[3]. For instance, every example of a problem we have given so far is Fortran-computable in this sense. Indeed, one might even imagine at this point that *every* problem is Fortran-computable.

The difficulty we now face is that there are many computer languages. That is, we also have, defined in a similar way, "C-computable", "Applesoft-computable", etc. But our goal here is to capture by a general definition an abstract notion of "computable" — i.e., to isolate the general structure of the computing process itself. The danger we face is that the various types of computability, as defined here, will say more about the individual languages that gave rise to them than they do about this general structure. But anyone who is familiar two or more languages will realize that this difficulty is more one of principle than one of practice. Consider two languages, e.g., Fortran and C. You can write a Fortran-emulator in C (and, indeed, this is probably what "Fortran" really is!). That is, you can write a C program that will accept as input lines of Fortran code: "Set x = 7", etc. The C program will then parse each such Fortran command, figure out what the Fortran language would have done to implement that command, and then itself do precisely that. From the mere existence of such a Fortran-emulator in C, it follows that every Fortran-computable problem is also C-computable. [To see this, consider any Fortran-computable problem π. Taking the Fortran program that computes this π and applying to it our emulator, we obtain a C program that computes π.] In a similar way, we can write a C-emulator in Fortran. We conclude, then, that the Fortran-computable problems are precisely the same as the C-computable problems. A similar argument shows that all the standard languages of the computer world generate precisely the same computable problems.

> **Exercise.** Consider three languages, A, B, and C. Suppose you were given an A-emulator in B, and a B-emulator in C. Could you use these two to construct an A-emulator in C? As a second question, consider two languages, A and B. Suppose that the A-

[3]To make this idea into a proper definition, we would have to specify the details of the language "Fortran". We shall not do this, since this entire discussion is intended merely as motivation for what follows.

computable problems are precisely the same as the B-computable problems. Can you exploit this fact to build an A-emulator in B?

How shall we turn this intuitive discussion into mathematics? Lest you imagine that this will be an easy exercise, we now introduce two new "languages".

The first, which we might call "MiniFortran", has just two commands: "Input", which allows the user to input any string S; and "Print $\emptyset$", which causes the computer to print the empty string. There is just one MiniFortran-computable problem, namely that with $\pi(S) = \emptyset$ for every input string S. Clearly, then, there are many fewer MiniFortran-computable problems than Fortran-computable problems. The problem with MiniFortran, of course, is that it is absurdly barren. A language must have a certain degree of richness [essentially, i) the ability to store plenty of intermediate data, ii) the ability to manipulate the data, and iii) the ability to branch, in response to those data] if it is to reach the mainstream (Fortran, C, etc) of computable problems.

Our second language, "HyperFortran", contains all the commands of Fortran, together with one additional command, with the following structure: "Do, for $I = 1, \infty\{\cdots\}$ Next". Here, "$\{\cdots\}$" consists of various Fortran commands, including those that may reset certain variables. The rule is that the computer will always exit from such a command (i.e., it will never hang here), and on doing so the variables will be set as follows. Consider any one variable, "x". If, in the course of the execution of this Do-loop, the variable "x" was set to some value, and did not ever change that assigned value beyond some particular iteration (i.e., beyond some particular I-value), then on exit from this command "x" is to be assigned that value. If, on the other hand, "x" changed its value an infinite number of times during the course of the Do-loop, then on exit "x" is assigned value $\emptyset$. This is, arguably, a legitimate computer command, at least in the sense that it is completely determined what is to be done in response to such a line of code. I grant that HyperFortran seems a little strange at first sight, but, absent a careful definition of the term "language", a reasonable case could, perhaps, be made that it is one. Note, incidentally, that in HyperFortran we can solve Goldbach's conjecture. We would use a program of the following form:

Do, for $I = 1, \infty$
If ($x == \emptyset$ and I is a Goldbach-counterexample) set $x = I$

Next
Print x

If Goldbach's conjecture is true, then there will be returned the empty string. If it is false, then there will be returned a counterexample to that conjecture.

We all know in our hearts that HyperFortran is unacceptable, but it is not so easy to spell out exactly why. True, you cannot run it on a physical computer – but you can't run Fortran, either; and in any case it is usually a bad idea to try to base mathematics on physical implementability. What HyperFortran does is illustrate that some care is going to be necessary in order to formulate a suitable definition of "computable".

So, to summarize, all "reasonable" computer languages (in some sense we have yet to pin down) seem to produce the same computable problems. Our challenge is to turn this intuitive idea — which is called *Church's Thesis* — into a piece of mathematics. There are at least three different strategies by which one might imagine doing this.

The first strategy begins by producing a formal definition of "reasonable language". This definition would be along the following lines. A "reasonable language" must have some commands. [Presumably, there would be just a finite number of types of commands, but, since arbitrary strings can normally appear in certain commands, there would be within these types an infinite number of actual commands. (Just like Fortran.)] With each command there would be associated something to "do" (such as manipulating a string, going somewhere, etc). We would require, as part of this definition, that these commands be rich enough to allow one to do "the necessary things for computing" (i.e., store arbitrary amounts of data, manipulate data, input/output, branch), but not so rich that they do "ridiculous things" (such as "Do, for $I = 1, \infty$"). Note that we are *not* specifying any specific language here, but rather are describing by the definition the characteristics that we will demand of a language in order that we deem it "reasonable". Then given such a language, $\mathcal{L}$, we would call a problem π *$\mathcal{L}$-computable* if there exists a program in $\mathcal{L}$ that, for every input string S, eventually halts, returning exactly $\pi(S)$. Finally, (the crowning result of this strategy) we would prove the following Theorem: For $\mathcal{L}$ and $\mathcal{L}'$ any two reasonable languages as defined above, the $\mathcal{L}$-computable problems are precisely the same as the $\mathcal{L}'$-computable problems. The key to this strategy, of course, is discovering the right definition: It has to look simple and not contrived, and at the same

time be just right so that the Theorem really is a theorem. I feel that it might be enlightening to carry out this strategy — but it looks like a lot of work.

The second strategy begins by noting that, since strings can be replaced by integers, each problem thereby becomes an integer-valued function on integers. We would now introduce some axioms that are intended to characterize the "computable functions". There might be a few simple ones, such as "Every constant function is computable."; and "The composition of two computable functions is computable." But then there would be some more complicated ones, requiring that certain constructions involving computable functions result in computable functions. Then, a function would be deemed computable if it arises from these axioms. This strategy has in fact been carried out: It is the subject called recursive function theory. [Recursive functions are precisely the computable (to be defined shortly) problems.] This strikes me as an elegant and promising approach. Its downside is that recursive function theory doesn't seem especially well-matched to physics — and, in particular, not to quantum mechanics. Furthermore, the subject of computational difficulty doesn't, as far as I am aware, fit in naturally with this strategy.

The third strategy consists of inventing the "simplest possible language" that is still (barely) rich enough that that it generates the same computable problems as the real-world languages. We then take computability to mean computability in this language. This is the strategy we shall pursue, in the following section.

5 Turing Machines

Fix a character set, $\mathcal{C}$. A Turing machine, operating with this character set, has two parts.

First, there is the machine itself. It is capable of being in any of a finite list of machine states, $q_1, q_2, \cdots, q_n, q_H$. Of these $(n+1)$ states, the last one, q_H, has a special role, as we shall see shortly. These machine states serve as the RAM: The machine will store data temporarily by the choice of the particular state in which it currently resides.

Second, there is a semi-infinite tape, divided into a succession of square boxes. Thus, at the beginning of the tape there appears the first box, fol-

lowed, moving along the tape, by the second, then the third, etc. There is no "final end" of the tape, i.e., there is available as much tape and as many boxes, so arranged, as might be needed. Each of these boxes may have printed within it a single character from the set $\mathcal{C}$, or the box may be blank (having no character). We denote this blank box-state by $\emptyset$ (not to be confused with the empty string). This tape serves as the hard drive: The machine will store data here on a more permanent basis for later use in the computation.

The machine also has a read/write head, which at any one moment resides over one of the squares of the tape. Thus, the complete state of this system (machine, tape, and head), at any moment, is characterized by specifying i) the internal state of the machine, ii) the characters printed on the tape, and iii) over which square the head currently resides. For example, a typical system-state might be: "machine state q_7; tape configuration '3 \$ v Z x $\emptyset$ $\emptyset$ $\emptyset$ $\cdots$'; head over the fourth square". In this configuration, the head would be read the character "Z"; and would print to the fourth square. It will turn out that the only tape-configurations of interest will be those in which all squares of the tape beyond a certain one are blank.

This machine operates by going from one system-state to the next according to certain rules that are set down in a table (the "program"). A typical row in this table is given below:

Curr State	Curr Char	$\rightarrow$	New State	New Char	Move
q_7	Z		q_3	p	L

This row demands that, if the computer finds itself in machine state q_7, with the head reading character "Z" on the tape, then the computer is to i) change its internal state to q_3, ii) erase the character "Z" from that square on the tape and print instead character "p", and iii) move the head one square to the left (i.e., toward the beginning of the tape). Thus, this particular row in the table would send the system from the system-state given in the previous paragraph to the following system-state: "machine state q_3; tape configuration '3 \$ v p x $\emptyset$ $\emptyset$ $\emptyset$ $\cdots$'; head over the third square". The full table for our Turing machine will consist of many such rows. In each row: The first entry must be one of the machine states $q_1, \cdots, q_n$ (but *not* the state q_H); the second entry must be a character, or the blank, $\emptyset$; the third entry

must be a machine state (with q_H allowed here); the fourth entry must be a character or the blank; and the fifth entry must be either "R" ("right") or "L" ("left"). Finally, the full table must contain one and only one such row for every possible choice of the first two entries. Thus, if there were ten machine states (including q_H), and the character set had six elements, then the full table would have exactly 63 ($= 9 \times 7$) rows. The Turing machine now operates in the obvious way. At each stage, it looks up its current machine-state and character-under-the-head in this table. It then reads off from the table what should be its next machine state, what the head should print on that square of the tape, and what movement the head should make (just one square, either to the right or to the left). If ever the machine finds itself in state q_H, then the machine halts (stops computing). That is why we do not allow the q_H-state as the first entry of any row.

The crucial features of this design are i) that the number of internal machines states is finite, while the number of tape-squares is infinite, and ii) what the machine will do next depends *only* on on the current machine-state and the character under the head, and not on what is printed elsewhere on the tape or whether the head resides over the fourth square or the nineteenth square.

To run a Turing machine, select the input string S and print it, one character at a time, at the beginning of the tape, leaving all the other tape-squares blank. Begin with the head over the first square and the machine in initial state q_1. Now let the machine run, step by step, for each step looking up in the table what to do next. If the machine, during the course of its running, never achieves the halt-state, q_H, then it will continue running forever. Well, that's life. If, however, it does eventually achieve q_H and halt, then we read the output string from the tape, starting from the first square and continuing until we reach the first blank square. Note that, during the running of every Turing machine, the tape always contains but a finite number of non-blank characters (although, of course, the number of such characters may, as S runs over all possible input strings, grow without bound). The crucial feature of this operation is that the table is to be fixed, once and for all, *before* we are given the input string S.

So, a Turing machine is a sort of mini-computer — a computer reduced to its essentials. First write the program. Then, input an initial string S. The computer computes away. Either it eventually halts, presenting an output string to you; or it runs forever, never presenting anything. A problem π over

a given character set is said to be *Turing-computable* if there exists a Turing machine (i.e., a choice of the number of internal states and of the table) that computes it (in particular, never halting, no matter what the input string S), as just described. To check whether you understand how a Turing machine works, try to convince yourself that you could write a Fortran-emulator of Turing. Assuming you have convinced yourself, then we may conclude (from the mere existence of such an emulator) that the Turing-computable problems is a subset of the Fortran-computable problems.

We give just one example of a Turing-computable problem. A string S is called a *palindrome* if it reads the same backwards as forwards, e.g., "$K9s4\$q\$4s9K$". Let the character set contain "y" and "n" (to make answers easier to express); and let the problem $\mathcal{S} \xrightarrow{\pi} \mathcal{S}$ be the following: $\pi(S)$ is "y" if S is a palindrome, and "n" if it is not. This problem is Turing-computable. We shall not write out the full table (which would have hundreds of rows!) that demonstrates this, but rather merely indicate how the machine would work.

The machine, in initial state q_1, reads the first entry in the string: Say it is a "K". The machine then goes into a state we call p_K, (whose description is "I've just read character 'K', and I'm now going to check to see if this is also the last character"), prints $\emptyset$, and moves one square to the right. If the head now reads anything other than $\emptyset$, the head moves another square to the right without changing anything. [That is, the table entry for "current state p_K and current character {nonblank}" requires remaining in p_K, reprinting whatever is already in that square on the tape, and then moving one square to the right.] The head thus continues moving to the right, one step at a time, until it encounters a blank square. On encountering a blank square, the machine goes into a new state we call r_K (whose description is "I'm now ready to compare the last character of the string with K"), reprints $\emptyset$, and moves one square to the left. The table entry for "current state r_K, and current character anything but K" puts the machine into a new state q_n (whose description is "This string is not a palindrome. Tough luck. I'm going to go back to the beginning now, to report that fact." We'll return later to how this is done.) The table entry for "current state r_K, and current character K" puts the machine into a state q_2 (whose description is "So far so good. I'll go back to the beginning now and get the next character."), prints $\emptyset$, and moves one square to the left. As long as the head continues to encounter nonblank squares, it continues to move leftward back over the string. [That is, the table

entry for "current state q_2, current character {nonblank}" retains the state q_2, reprints whatever is already under the head, and moves one square to the left.] However, as soon as the head meets a blank square, the machine goes back to state q_1, prints $\emptyset$, and moves one square to the right. The process now starts over (but now with a shorter string, for we have just removed from the original string S its first and last characters). That is, the machine reads the current square (yielding, say, character "9") goes into state p_9, moves to the right until it encounters a blank square, goes into state r_9, carries out a comparison of "9" with the current character, goes into either state q_n or q_s, etc. Continue in this way. If, eventually, the string is exhausted, then the machine goes into state q_y (whose description is "It is a palindrome! I can't wait to to back to the beginning and deliver the good news"). [That is, the table entry for "current state q_1, current character blank" places the machine in state q_y, prints $\emptyset$, and moves one square to the left.]

The process above eventually places the machine in either the state q_n or the state q_y. How does the reporting of the news ("y" or "n") work? We want the machine states q_n and q_y to move the head to the left, for that is where the reporting must take place. But how will our Turing machine know when it has reached the leftmost square? [There will just be blanks back there, for we have now erased the initial portions of our original string S.] One way to accomplish this is to move, initially, the entire string S up the tape a little bit, to make room for a marker at square one. Here is how to move the entire string S one square to the right. Read the first character (say, "K", again). Go to state s_K (whose description is "I'm about to move a 'K' one square to the right"), and move the head one square to the right. If, say, the next square read by the head contains the character "9" then print the "K" in this square, go to state s_9, and move another square to the right. [That is, these are the instructions for "current machine state s_K, current character 9".] Continue until you reach a blank square (i.e., to the end of the string S). Then print the last character (as determined by what s-state you happen to be in at the time), and go into a state that causes movement to the left until you encounter a $\emptyset$. You have now moved the entire initial string S one square to the right. In a similar way, we may move the initial string to the right a second square, and print anything (say, "v") in the first square of the tape. All of this would be done *before* the program of the previous paragraph. Do this, and then run that program (on the original string S, as now displaced).

We next describe how the reporting works. The state q_y will require motion of the head to the left continue as long as that head encounters only blank squares. But as soon as it encounters character "v" on the tape (i.e., as soon as it reaches the first square of the tape), it will print "y" and go to the halt state, q_H. In this way there is returned that the original string S was a palindrome. The state q_n has to work a little differently. It will cause the head to continue moving to the left as long as there is encountered nonblank characters. But, according to q_n, as soon as the head encounters a blank, the machine goes into still another another state, q_{nn} (whose description is "OK. Now all I need is to is find that "v" off to my left, to whom I must my report"). So, the table entry for "current state q_{nn}, current character v" is "Go to state q_H, print character 'n', and move one square to the right." In this way there is reported that the original string S was not a palindrome.

Well, that was exhausting. Suppose our initial character set contained m characters. Then we must introduce $3m$ machine states, for the p's, r's, and s; as well as additional nine machine states, for the various q's, including q_H. Thus, there will be a total of $(3m + 8)(m + 1)$ rows in the table. Even for just ten characters, for instance, this is a total of 418 rows! You might think it would have been easier to have the machine simply remember the string S as it passes over it the first time, and then just make a single check for palindrome-ness when it reaches the far end of the string. But that won't work, because the machine is allowed only a finite number of internal states, and this number must be fixed already in the original table — and there is no adjusting that number depending on the string S.

> **Exercise.** Why do we not, in our definition of a Turing machine, fix the total number of machine states, once and for all? Because that doesn't work. Fix the character set, $\mathcal{C}$. Prove that, for every integer n, there exists a problem (over that character set) that cannot be solved by any Turing machine with just $(n+1)$ internal states.

> **Exercise.** Let the character set be any one that includes the ten digits together with "n". Convince yourself that you could build a Turing machine that returns "n" if input string S is not an integer; and $S + 1$ if it is an integer. Convince yourself that you could build a Turing machine that, whenever the string S

is two digits separated by a single "n", returns their sum; and otherwise returns "n". Convince yourself that you could build a Turing machine that not only solves the problem of the last paragraph, but cleans up the tape (i.e., removes everything but the answer-string) before reporting.

The next step in learning this subject is to play with Turing machines so as to get a feeling for what they can do. Convince yourself that you could build machines (but don't actually do it!) to solve successively harder problems. Start with easy problems, such as those of the exercise above. Then try harder ones: multiplication of integers, division of integers with remainder, deciding whether or not an integer is prime, deciding whether or not an integer is a counterexample to the Goldbach conjecture, etc. Through this process, you must eventually convince yourself of the following key fact: **There exists a Fortran-emulator in Turing.** [See, e.g., [**?**] for some details.] As a consequence of this fact, the Turing-computable problems are the same as the Fortran-computable problems; and, by similar arguments, as the C-computable problems, as the Applesoft-computable problems, etc. The idea, then, is that the Turing-language is the simplest one that still has sufficient richness that it generates the "right" computable problems.

Here is our main definition: A problem π, over a given character set $\mathcal{C}$, is said to be *computable* if it is Turing-computable — that is, if there exists a Turing computer T that, run on any string $S \in \mathcal{S}$, always eventually halts, returning $\pi(S)$. What you have done in the paragraph above should convince you that this is a reasonable definition. What most people do in this subject, I believe, is "talk in terms of Turing machines, but think in terms of their favorite language (whatever it happens to be)". We emphasize that, while the psychological situation here is complex, the mathematical one is not: We define a Turing machine; and, using it, we define a computable problem.

Every problem we have discussed so far is computable (including the one that sends every string to "y" if Goldbach's conjecture is true; and every string to "n" if that conjecture is false). The composition of two computable problems is computable. [This fact is useful in showing that suitable changes in the input/output grammar do not affect computability.] For π and π' computable problems, the problem π'' with $\pi''(S) = \pi(S)\ \pi'(S)$ (concatenation of strings on the right) is computable.

Many constructions involving Turing machines rest on the following fact:

Every Turing machine can be represented as a string. Here is one way to do this. Consider a Turing machine T over character set $\mathcal{C}$. Let, for example, the first few rows of the table for T be:

Curr State	Curr Char	$\rightarrow$	New State	New Char	Move
q_7	Z		q_H	p	L
q_{11}	$\emptyset$		q_7	\$	R
q_8	2		q_8	Z	R

The first step in rewriting this T as a string is to introduce the new character set, $\mathcal{C}'$, that results from adding one additional character, say "$*$", to $\mathcal{C}$. [We are assuming here that "$*$" does not denote any element of the original character set $\mathcal{C}$ itself. This element "$*$" will serve as a marker. More on this later.] The next step is to choose a string over $\mathcal{C}$ to represent each machine state for the Turing machine T. For example, we might represent states q_7, q_H, q_{11}, and q_8 by strings "$s6$", "$\$B4$", "$uU$", and "8", respectively. Then the rows of the table above would be represented by a string as follows:

$$*s6 * Z * \$B4 * p * *uU * *s6 * \$ * * * 8 * 2 * 8 * Z * * * \cdots \qquad (1)$$

We have simply written the entries in the table (replacing each machine state by its string), row by row, one after another, using the "$*$" to separate the entries. The reading or writing of a blank square is indicated by placing nothing between the two separators: "$**$". Movement of the head to the left is indicated by placing nothing between the separators ("$**$"); to the right by a "$*$" between them ("$* * *$").

Thus, each Turing machine over $\mathcal{C}$ gives rise to some string over $\mathcal{C}' = \mathcal{C} \cup \{*\}$. The machine for the palindrome problem with ten characters, for example, results in a string of about 4,800 characters. Note that a given Turing machine can be represented by a string in many ways — e.g., by choosing different strings to represent the various machine states, or by changing the order in which the rows of the table are presented.

The time has come to simplify our language a little bit. In Sect. 3, we introduced a problem π, on character set $\{0, 1, \cdots, 9, f, y, n\}$; with $\pi(S)$ equal to "f", "y", or "n" according as S is not an integer greater than equal to two, is a prime integer, or is a nonprime integer. We shall now allow ourselves to describe this as "the problem of deciding whether or not an integer

is prime". Thus, in this description, it is understood that i) the character set has sufficient characters to describe the input strings of interest (i.e., here, the digits), ii) any strings constructed from those or other characters, that are not the strings of interest (e.g., here, "007"), will be suitably branded by the problem (e.g., sent to "f"), and iii) the outcomes of interest (here, "prime" and "not prime") will be suitably encoded as strings over our character set. We can safely ignore how such details are arranged, and thereby avoid an unnecessary distraction. Next, recall, from the previous paragraph, that a Turing machine over character set $\mathcal{C}$ can be represented as a string over the character set $\mathcal{C}' = \mathcal{C} \cup \{*\}$, the extra character "$*$" having been introduced as a marker. Now fix any ordering for $\mathcal{C}'$, thus obtaining, as we noted earlier, an ordering (dictionary) for the strings over $\mathcal{C}'$, and thereby an assignment of an integer to each such string. Combining these two constructions, then, we assign, to each Turing machine over $\mathcal{C}$, an integer (although, of course, not every integer arises from some Turing machine). Here is a somewhat more useful assignment. Consider the first string over $\mathcal{C}'$ (in this ordering) that represents a Turing machine, and call that machine number one; then consider the second string that represents a Turing machine, and call that machine number two; then the third; etc. In this way, we assign to each Turing machine an integer, such that now each integer also represents some Turing machine. Thus, we may speak of "the n^{th} Turing machine", implicitly invoking this numbering. Next, we may combine this construction with our correspondence between strings over (the now ordered) $\mathcal{C}$ and integers. There results an assignment, to each Turing machine over $\mathcal{C}$, of a string over this same character set; such that each string now represents some Turing machine. We denote by T_S the Turing machine associated with string S. Shortly, we will want to turn a pair, such as (T, S), where T is a Turing machine over $\mathcal{C}$ and S a string over $\mathcal{C}$, into a single string over $\mathcal{C}$. We may do this, e.g., in the following manner. First take the string over $\mathcal{C}' = \mathcal{C} \cup \{*\}$ that represents T (as above), then append "$*****$" (a marker, to separate the representation of T from S), and finally append the string S. In this way, we represent (T, S) as a string over $\mathcal{C}'$. But now we may convert this to an integer — or to a string over $\mathcal{C}$ — using the techniques above. If you find yourself uncomfortable with all these conventions, you might try to restore the missing material for a short while, until you get used to them.

Exercise. Convince yourself that each of the following problems

is computable: i) that of deciding whether or not a string over $\mathcal{C}'$ represents *some* Turing machine; ii) that of deciding whether or not two strings represent the same Turing machine (where by "same" we mean "differing only in rearrangement of the machine states (preserving q_1 and q_H), and in the order in which the rows are presented in the table"); iii) that which sends integer S to the string for the S^{th} Turing machine; iv) that which sends integer S to the string for the S^{th} Turing machine, eliminating repetitions (via "same"); v) any problem π such that $\pi(S) = \emptyset$ for all but at most a finite number of strings vi) the problem that assigns, to each string S, the positive integer that is the number of steps Turing machine T takes, on string S, before it halts; where T is some fixed Turing machine that does halt for every input string. Much more difficult, e.g., is the problem of deciding whether two Turing machines compute the same problem (or, indeed, whether a given Turing machine T computes any problem at all, i.e., whether that machine always halts, no matter what the input string).

6 Noncomputable Problems

It is not hard to convince yourself that every problem is computable. A problem, after all, is merely a mapping $\mathcal{S} \stackrel{\pi}{\rightarrow} \mathcal{S}$. So, to specify a problem, you must specify what the mapping is; i.e., specify how to determine, for any string $S \in \mathcal{S}$, some string, $\pi(S)$; i.e., specify how to compute, given any S, some $\pi(S)$. But "compute", we've come to realize, means "Turing-compute".

But, while this intuitive argument may seem plausible, it is simply wrong: There do indeed exist non-computable problems. The easiest way to prove this is by a cardinality argument. The set of all Turing machines that compute problems is countable (since it is a subset of the set of all Turing machines; which in turn can be represented as a subset of the (countable) set of strings over some character set). But the set of all problems, as we saw in Sect. 3, is uncountably infinite. Therefore, the mapping "send machine to the problem it computes" from the former to the latter cannot be onto.

While the above proof is simple, it doesn't give much insight into which problems are noncomputable and which are not. Fortunately, it turns out that there is an example that is both simple and illuminating.

> The **halting problem** is that mapping $\mathcal{S} \stackrel{\pi}{\to} \mathcal{S}$ that sends Turing machine T and string S to "halt" if the machine T, running on input string S, eventually halts, and to "not halt" if that machine on that string continues running indefinitely without halting.

Note that the halting problem is indeed a problem, for, given machine T and string S, then T on S either halts, or it does not. You might imagine that we could build a master Turing machine, **H**, that would compute the halting problem, in the following manner: Given (T, S), where T is some Turing machine and S some string, **H** would merely simulate the action of T on S, doing what T would do, step by step, and ultimately reporting the result: "halt" or "not halt". But, unfortunately, this doesn't work. There is no difficulty if T, applied to S, ultimately halts. Then **H** will discover this eventually, and duly report "halt". But what if T, applied to S, never halts? There will in this case never be a moment when **H** discovers this fact; and so no moment when **H** will report "not halt".

Now comes the central result of this subject:

Theorem. The halting problem is not computable.

Proof: For S any string, denote by T_S the Turing machine represented by that string, as described above. Suppose, for contradiction, that there existed a Turing machine, **H**, that computes the halting problem, reporting $\mathbf{H}(T, S) =$ "halt" or $\mathbf{H}(T, S) =$ "not halt", according as machine T, applied to string S, halts or not. We now construct a new Turing machine, $\tilde{T}$, as follows. Given string S, $\tilde{T}$ first runs machine **H** on (T_S, S), and then proceeds as follows; If $\mathbf{H}(T_S, S) =$ "halt", then $\tilde{T}$ continues running, without ever halting; while if $\mathbf{H}(T_S, S) =$ "not halt", then $\tilde{T}$ immediately halts. [In other words, we build a Turing machine $\tilde{T}$ that, given string S, asks **H** about (T_S, S), and then *does* the opposite of what **H** *reports*.] Now, $\tilde{T}$ is a Turing machine, and so it is represented by some string: $\tilde{T} = T_{\tilde{S}}$, for some $\tilde{S}$. We now ask: What happens when machine $\tilde{T}$ is run on string $\tilde{S}$? Suppose, say, that it eventually halts. But this means, from the way we defined machine $\tilde{T}$, that $\mathbf{H}(T_{\tilde{S}}, \tilde{S}) =$ "not halt". But this means, from the defining property of **H** that machine $T_{\tilde{S}}$ $(= \tilde{T})$, when run on string $\tilde{S}$, does not halt. This is a contradiction. Similarly, the supposition that machine $\tilde{T}$, run on string $\tilde{S}$, does not halt leads to a contradiction. We thus conclude, since the assumption that there

exists a Turing machine **H** that computes the halting problem leads to a contradiction, that the halting problem is not computable. \

This proof — essentially, a diagonal argument — is at the same time very simple and very confusing. I urge you to return to it in the coming weeks, as often as necessary, until you have mastered it. The discussion below is intended to give you a feeling for what the theorem means.

Note that the theorem does *not* assert that there is a *specific* machine T and string S such that we will be unable to decide whether that T, run on that S, halts. Indeed, we expect that, given (T, S), we could, given enough time and ingenuity, determine whether halting occurs. What the theorem *does* assert is that there is no single algorithm that will correctly decide halting in *every* case, i.e., for *every* (T, S).

Here is a more poignant restatement of the paragraph above. Imagine having the following job: Occasionally, there is brought to you a Turing machine, T, and string, S, and you are to determine and report to your boss whether or not that machine, applied to that string, ever halts. In some cases — e.g., a machine for which q_H never appears in the third column of the table; or for which all states in the third column are q_H — your decision will take but a few minutes. In other cases — e.g., that in which there is a collection of machine states i) from which the machine cannot exit, ii) such that q_H does not appear in the third column for any of these states, and iii) into which the machine, by virtue of the given S, will enter — it may take take hours. In more complicated cases it might take days . . . or even years. As you continue working in this job, you will build a repertoire of arguments for settling this question in specific cases. And you will note that you are continually adding new, ever more clever, arguments to your collection. At some point, you may ask yourself: "Will this job ever become routine? Will I ever reach the point at which I have developed all the arguments that are needed to solve these puzzles — the point at which no further originality will be required for this job?" These questions are answered by the theorem above: The answers are all "No".

Suppose for a moment that we had felt inclined to include use of the additional command "Do, for $I = 1, \infty$ $\{\cdots\}$ Next" (i.e., the infinite Do-loop) in our notion of "computable". As a result, as we have noted, there would be more computable problems. However, we could still define the halting problem (now referring to Turing machines in which this additional command is allowed). But the theorem above would still hold in this case

(for its proof would go through in the same way). In other words, we would conclude that, even in this stronger language, we cannot compute the halting problem for that language.

Next, suppose for a moment that we had a master Turing machine **H** that did compute the halting problem. Then, we claim, we could resolve the Goldbach conjecture. We do this as follows. Construct a Turing machine T that, applied to any string S, ignores S completely, and starts searching the even integers $(4, 6, \cdots)$ looking for a Goldbach-counterexample. If it finds a counterexample, it halts, announcing this result. As long as T hasn't yet found a counterexample, it just keeps looking. Now, all we have to do, having built this machine T, is run the master machine **H** on machine T and any string S. If the result is $\mathbf{H}(T, S) =$ "halt", then the Goldbach conjecture is false; if "not halt", true. Note that we settle this conjecture without doing any real work: We don't have to have deep thoughts about the structure of the primes, or about any other relevant mathematics. All we need in order to resolve the conjecture is, essentially, an understanding of what it is asking for. In a similar way, we could use **H** to resolve, again without doing any real work, many of the other open questions in mathematics. In short, a great deal of mathematics can be encoded into the question of whether certain Turing machines halt. Perhaps this observation makes the theorem seem less surprising.

One occasionally reads, in the Sunday supplement, an article suggesting that physics is dead: that we have now discovered the fundamental structure of Nature — the "theory of everything" — and that all that remains is working out the details. Of course, this is a mere guess on the part of the writer: Nobody has (or can have?) any real insight into this question. But note that mathematics is very different from physics in this regard. Mathematics isn't dead yet; and, we suggest, it never will be. Indeed, we have a *theorem* to the effect that new and different insights will always be required in the development of mathematics!

> **Exercise.** Show that the following problems are not computable: i) the problem that asks whether a given Turing machine solves *some* problem; ii) the problem that asks whether, given a Turing machine, there is *some* string S on which it halts; iii) the problem of deciding whether two Turing machines (both of which do solve some problem) solve the same problem. [Hint: Show that a Tur-

ing machine that computes these problems could be reconfigured to give a Turing machine that computes the halting problem.]

This paragraph is mere whimsy, which you should feel free to ignore. I would like to suggest that the expression "X never happens" (as well as its various siblings) has no real meaning whatsoever. Rather, this expression is merely a part of a sociological convention: We have all agreed that, when we hear it, we shall nod our heads knowingly (rather than, say, rolling our eyes). Certainly, this idea is not necessary to function in our daily lives. Wolves, for example, neither use nor understand this expression, and yet they get along, in the woods, quite well. Imagine an individual who has been raised by wolves, and shares *their* sociology. You wish to explain to this person that "This Turing machine, when run on this string, never halts." This individual replies "I have no idea what you are talking about." You say "Well, the machine doesn't halt after 9 steps." "Right." "It doesn't halt after 137 steps." "Right." "And, in fact, it doesn't halt after any number of steps." "I have no idea what you are talking about." Or, you might try to argue using the structure of a particular Turing machine T. "The state q_H nowhere appears as the third entry in any row." "Right." "Therefore, the machine doesn't halt after 19 steps, because it couldn't be in the state q_H then." "Right." "Similarly, it doesn't halt after n steps, for any $n = 1, 2, \cdots$." "I have no idea what you are talking about." Your growing sense of frustration arises from your inability to express this idea in terms of anything else. Here is another perspective. Imagine that you were transported to another planet, the residents of which have decided to explain to you a term, "swerm", in their language. You find yourself on the other side of the conversation. They say "Horses are brown." "Right." "And three is an integer." "Right." "And swerm." "I have no idea what you are talking about." They try to argue in more detail — exploring the light reflection from horse's coats, and counting up to three — but still you cannot understand. They feel a growing frustration with your skepticism. The residents of this planet, it turns out, have introduced the notion of a Turing machine, which they may use to compute problems. Particularly famous is the **swerm problem**: Given string S let $\pi(S)$ be "swerm" or "not swerm" according to whether or not the string S is swerm. [You, of course, suspect that this is not a problem at all, but just nonsense-talk.] In fact, they have even proven what they regard as an important theorem: The swerm problem is not computable. "At last.", you

think, "Surely now, by merely going through their proof, I will be able to understand what this 'swerm' is all about." So, you go to their library, find the paper containing this proof, and begin to read. But you quickly discover, to your dismay, that their so-called "proof" tells you nothing whatever – for it uses, in an essential way, the very concept of swerm. [Apparently, the people of this planet have decided that it is appropriate to use swerm within proofs.] You conclude that, at worst, the residents of this planet are delusional. At best, they have managed to discover that it is not possible to construct a Turing machines that will explicate their strange sociology. As an exercise, try rereading the previous several pages, mentally substituting, everywhere, "swerm" for "halt".

7 Noncomputable Numbers

As an example of an application of Turing machines, we now consider briefly the subject of noncomputable numbers.

A positive real number x is said to be *computable* if there exists a Turing machine that, when applied to any positive integer S as input, returns some rational number, a/b, such that $|x - a/b| \leq 1/S$. In other words, the computable numbers are those to which we may compute approximations. Note that the two integers a and b in the fraction must be encoded into a single string in the output (e.g., by using a separator, and then translating back to the original character set). The reason that we approximate x by rationals is that it is easy to express a rational number in terms of a string. Note also that many Turing machines may compute the same number x (e.g., by providing different rational approximations to it). And finally, note that the function "$1/S$" on the right of the equation above could as well be replaced by any (computable) function of S that decreases monotonically to zero, e.g., $1/S^7$, or e^{-S}, resulting in the same notion of computable number: You can easily retrofit a Turing machine designed for one function on the right to one designed for another. The problem of whether a given Turing machine "computes" some real number x in this sense; as well as that of deciding whether two machines compute the same number, is not computable.

> *Exercise.* Call a number x *hypercomputable* if there exists a Turing machine that, given integer S as input, returns a Turing machine that computes (in the sense above) some real number y such

that $|x - y| \leq 1/S$. Clearly, every computable number is hypercomputable. Is every hypercomputable number computable?

The number e, for example, is computable. An appropriate Turing machine might use the formula $e = 1 + 1/1! + 1/2! + \cdots$, keeping enough terms and computing the terms with sufficient accuracy to determine a rational within $1/S$ of e. Similarly, the number $\log(\sin^{-1}(.714) + \sinh(e/4))/\pi^{2.7}$ is computable, as is every other other number you might think of offhand. Note that whether or not a number is computable depends only on the **number** itself, and not how that number is expressed. Thus, every rational number is computable, as is the number that is "1" if Goldbach's conjecture is true; and "0" if it is false. Indeed, it is tempting to imagine that every number might be computable. But, there do indeed exist noncomputable numbers, as follows immediately by a cardinality argument: The set of real numbers is uncountably infinite, while the set of Turing machines is only countable infinite.

Again, as with the case of noncomputable problems, we would like, not merely an existence argument, but a "concrete" example. Here is one. Set

$$c = \sum_{n=1}^{\infty} a_n/3^n, \tag{2}$$

where $a_n = 2$ if, for the n^{th} pair (T, S), the result of running the Turing machine T on the string S halts; and $a_n = 0$ if that machine on that string does not halt. Note that, since each Turing machine T on each string S either halts or does not halt, each a_n is a definite number; and so this c is also a perfectly definite number. If you know this c to sufficient accuracy, then you know whether each of the first n Turing machines/strings halts. Indeed, either $c < 1/3$ (the case in which the first Turing machine/string does not halt), or $c > 2/3$ (the case in which it does halt); so knowing c within $(1/6)$ determines whether or not the first machine/string halts. Similarly, knowing c to within $1/(2 * 3^n)$ determines whether each of the first n machine/strings halt. It follows from these remarks that the number c is not computable. Indeed, suppose we were given Turing machine, $\tilde{T}$, that computes c, in the sense described above. Then, we could easily rebuild that machine into one that computes the halting problem, as follows: If you wish to know whether the n^{th} pair (T, S) halts, apply this $\tilde{T}$ to string $2 * 3^n$ (written out as its

digits), and interpret the rational number that results. But we know that the halting problem is not computable, and so no such machine $\tilde{T}$ exists. Here is a curious corollary of these observations: The number c above is not rational. Note that this is not at all obvious from the formula (**??**).

> **Exercise.** Show that there exists a Turing machine that accepts as input a positive integer S, returning a rational a/b, such that the resulting sequence of rational numbers increases monotonically and converges to c from below. [Hint: Given n, run, for each $k < n$, the k^{th} (T, S)-pair for $(n - k)$ steps.] Prove that if number x is such that there exists a Turing-generated monotonic sequence of rationals converging to it from below (in the sense of the previous sentence), and also one from above, then x is computable. Does there exist a non-computable number such that there exists neither such a sequence from above nor one from below? What about the number that results from (**??**) by replacing 3 by -3 on the right?

It is interesting to speculate what might happen if ever a physical theory were to predict, for the outcome of some experiment, a noncomputable number, e.g., the c above. Then, since c really is a number, the theory would be making a perfectly definite prediction for the outcome of the experiment. However, to evaluate that predicted number, to higher and higher precision, would require new and ever more sophisticated insights (for that is the meaning of not being computable). Thus, we might some day reach the situation in which the experimentalists, who have carried out the experiment to, say, one part in 10^7, are way ahead of the theoreticians, who have only been able to carry out the computation of what the theory predicts to one part in 10^2! And there would be no guarantee that any greater precision would be forthcoming from the theoreticians any time soon. [A more accurate determination of the prediction of the theory might require, for example, that Goldbach's conjecture be settled.] This speculation is not entirely idle, for there are some (very weak) indications that noncomputable numbers may actually arise in some future quantum theory of gravity.

8 Formal Mathematics

The most famous application of computability is to a certain program for formalizing mathematics. We here merely touch on a few highlights of this subject: For more details, see, e.g., ([**?**]). Nothing in this section will be used later, so it may be skipped.

The idea is to apply the notions of the previous sections — strings, problems, computability, etc — to mathematics itself. But this program immediately runs into a serious roadblock: The ingredients of ordinary mathematics — the definitions, theorems and proofs — are informal in character. Here is an example.

> **Theorem.** There is no largest prime integer.
> Proof: Let, for contradiction, n be the largest prime integer. Set $a = n! + 1$. Factor this integer a as a product primes, and let p denote one of those prime factors. Then p cannot be 2, for the formula above shows that a is an even integer (namely, $n!$) plus 1, i.e., that the division of a by 2 leaves a remainder of 1. And, similarly, p cannot be 3, for the division of a by 3 also leaves a remainder of 1. So on, up to n. We conclude that $p > n$. Thus, this p is a prime number greater than n, which contradicts our choice of n.

What appears above is merely eight English sentences, so designed to cause the reader to nod in agreement. True, it is also a string (over some character set), but these are not the sorts of strings that can easily be manipulated by Turing machines, or about which we can easily prove theorems.

The idea of formal mathematics is to introduce a certain class of particularly simple strings, together with certain manipulations of those strings, which will reflect the content of ordinary mathematics. Thus, for example, certain strings will be deemed "assertions". Of course, these strings will be merely meaningless lists of symbols: They won't actually "assert" anything. And, similarly, other strings will be designated "proofs of assertions" (although they will not actually "prove" anything). This framework opens up the possibility of applying mathematics to mathematics itself, i.e., of proving (informal) theorems about the assertion-strings and proof-strings. We

further arrange matters so that the problem of deciding whether a given string is an assertion-string, or whether it is a proof-string, is computable. Thus, ultimately, we will be able to apply Turing machines and the ideas of computability to mathematics itself.

Mathematics is, essentially, set theory, and thus our goal is to formalize (i.e., render as strings) the subject of set theory. We emphasize again that mere string-manipulation should not be confused with mathematical "Truth". Think of "formal mathematics" as just another area of mathematics, analogous, e.g., to group theory — instead of manipulating group-elements according to the rules laid down for group theory, we shall manipulate various mathematics-strings according to some other set of rules we shall lay down. Adopting this perspective is more easily said than done.

Fix a character set, $\mathcal{C}$ (e.g., the set of lower-case Latin letters). We next introduce a new character set $\tilde{\mathcal{C}}$, consisting of the characters in $\mathcal{C}$, together with the following ten additional characters: $=$, $\in$, $\neg$, $\wedge$, $\forall$, $\}$, $\{$, $:$, $)$, and $($. Next, we introduce a certain collection of strings over $\tilde{\mathcal{C}}$, called the *formulae*. The rules are the following: i) For x and y any nonempty strings over $\mathcal{C}$, each of "$x = y$" and "$x \in y$" is a formula. ii) For $\mathcal{A}$ and $\mathcal{B}$ any formulae, each of "$\neg\mathcal{A}$" and "$(\mathcal{A} \wedge \mathcal{B})$" is a formula. iii) For $\mathcal{A}$ any formula, and x any nonempty $\mathcal{C}$-string, "$\forall x(\mathcal{A})$" is a formula. iv) The two expressions in item i) also result in formulae if either or both of x and y is instead replaced by a $\tilde{\mathcal{C}}$-string of the form "$\{z : \mathcal{A}\}$", where z is any nonempty $\mathcal{C}$-string and $\mathcal{A}$ is any formula. Using these rules, we may generate an enormous number of formulae, e.g., "$\forall x((y \in x \wedge \neg\forall s(z = y)) \wedge z \in \{w : x \in w\})$". A crucial fact about this construction is this: The problem of deciding whether or not a $\tilde{C}$-string is a formula is computable.

The nonempty strings over $\mathcal{C}$ are called *classes* (which we think of as "sets", the name having been changed for certain technical reasons). We also give these new symbols suggestive names: "$=$" is called "equals"; "$\in$" is called "is an element of"; "$\neg$" is called "not"; "$\wedge$" is called "and"; "$\forall$" is called "for all"; and "$\{z : \mathcal{A}\}$" is called "the collection of all sets z such that $\mathcal{A}$". The purpose of these names is merely to make the strings easier to remember and to think about: These names are not to be construed as bestowing any "meaning".

A *definition* is merely a shorthand way of a writing certain, commonly occurring, $\tilde{\mathcal{C}}$-strings. Here are a few examples of useful definitions (and their names): "$\mathcal{A} \vee \mathcal{B}$" stands for "$\neg(\neg\mathcal{A} \wedge \neg\mathcal{B})$" ("or"); "$\mathcal{A} \Rightarrow \mathcal{B}$" stands for

"$\neg\mathcal{A} \vee \mathcal{B}$" ("implies"); "$\exists x(\mathcal{A})$" stands for "$\neg\forall x(\neg\mathcal{A})$" ("there exists an x such that"); "$x \cup y$" stands for "$\{z : z \in x \vee z \in y\}$" ("union"); "$x \subset y$" stands for "$\forall z(z \in x \Rightarrow z \in y)$" ("subset"); "$\emptyset$" stands for "$\{z : \neg z = z\}$" ("empty set"); "$\{x\}$" stands for "$\{z : z = x\}$" ("set whose only element is x"); The integers are now defined as follows: 0 is defined as $\emptyset$; 1 as $0 \cup \{0\}$; 2 as $1 \cup \{1\}$; etc. Thus, for example, 5 denotes the set with precisely the following five elements: $0, 1, 2, 3$, and 4. There is also a definition (which we shall not give) of a set ω that deserves to be called the *integers*. Note that we cannot, e.g., merely write "$\omega = \{0, 1, 2, \cdots\}$", for neither "," nor "$\cdots$" are allowed symbols. We emphasize that these various definitions add nothing whatever to the logical structure (nor are we incorporating their symbols into our character set): Their only role is to make it easier to write certain long strings.

You probably once learned a version of set theory in which one begins with some basic universal set, e.g., "the set of all dogs"; then introduces various subsets of this basic set, e.g., "the set of all brown dogs"; and finally introduces the various set-relations — subsets, unions, etc. — on these subsets. Here, things are structured a little differently. Think of the class "x" (or any nonempty C-string) as having elements (just like a set), i.e., we might have "$y \in x$". But, on the other hand, this x might also be a member of some "z", i.e., we might have "$x \in z$". There is no basic "universal set", fixed at the beginning, from which flows all the other sets. All we have, instead, is this abstract hierarchy, running indefinitely in both directions: Classes are elements of classes that are themselves elements of classes, etc; and classes have elements that are themselves classes that have elements, etc.

You must now convince yourself that every assertion you have ever made — or are likely to make — in mathematics can be translated into a corresponding formula, as defined above. This exercise is similar to that of convincing yourself that everything you would have called a "procedure" or "algorithm" can be translated into a corresponding Turing machine. The following remarks are intended to get you started in this process. A rational number is defined, as usual, as a certain ordered pair of integers; and we may introduce the set of all rationals. A real number is defined as a certain set of rationals (namely, all rationals less than that number); and we may introduce the set of all reals. Thus, for example, the English sentence "x is a real number" is translated into a certain formula (i.e., into a certain string over $\tilde{C}$). The arithmetic operations on real numbers are defined as

corresponding manipulations of these sets. A mapping from set x to set y is defined as a set of ordered pairs, (a, b), with $a \in x$ and $b \in y$, such that every element of x is included once and only once as the first entry of one of these pairs. We now can introduce, for example, the real (or complex) functions of one (or more) real variables. Continuity and smoothness of such functions ("for every positive number ϵ there exists a number δ such that ...") is translated directly into the language of $\tilde{C}$-strings. Thus, for example, the English sentence "f is a continuous real-valued function of one real variable" is translated into a certain formula. Now consider some mathematical assertion, e.g., "Every smooth vector field on the 2-sphere vanishes somewhere." This assertion will be translated into a certain formula. Within this formula will be the definition of a 2-sphere (ordered triples, (x, y, z) of real numbers satisfying $x^2 + y^2 + z^2 = 1$) and of a vector field (a certain map sending each point of this 2-sphere to a tangent vector at that point). This formula will further include (using the symbol "$\wedge$") the condition that this vector field be smooth. Then, after all this, there will appear in our formula "$\Rightarrow$" (actually, the $\tilde{C}$-string this represents). And then, finally, there will appear the translation of "there exists a point of this 2-sphere at which that vector field vanishes". I urge you to play around with this, and other, examples until you have convinced yourself that the formulae, i.e., the $\tilde{C}$-strings specified above, are rich enough to encompass the language of mathematical objects and of mathematical assertions about those objects.

The next step is to isolate a certain collection of formulae, called the *axioms*. We shall not attempt to write out any axiom system (of which there are several) in detail, but rather merely indicate what those systems look like. Typical axioms might include "$\neg\neg\mathcal{A} \Rightarrow \mathcal{A}$" and "$\mathcal{A} \wedge \mathcal{B} \Rightarrow \mathcal{A}$" (logical axioms); "$x = y \Rightarrow \forall z(z \in x \Rightarrow z \in y)$" and "$\forall z(z \in \{z : \mathcal{A}\} \Rightarrow \mathcal{A})$" (tying "$=$" and "$\{z : \cdots\}$ in with "$\in$"); "$\exists y((x \in y \Rightarrow \exists w(x \in w \wedge w \in z)) \wedge (\exists w(x \in w \wedge w \in z) \Rightarrow x \in y))$" (existence of infinite unions); "$\forall x((\neg x = \emptyset) \Rightarrow (\exists y(y \in x \wedge x \cap y = \emptyset)))$" (which will, among other things guarantee that no class is an element of itself) and "$\exists x(\exists y(\emptyset \in x \wedge \forall z(z \in x \Rightarrow z \cup \{z\} \in x) \wedge x \in y))$" (which will, essentially, guarantee the existence of "infinite sets"). Other candidates for axioms might include formulae that reflect the axiom of choice, the axiom of the excluded middle, the axiom that every subset of $[0, 1]$ is measurable, etc. The crucial thing about these axiom systems is that they are so constructed that the problem of deciding whether or not a formula is an axiom is computable.

So, fix a suitable system of axioms. A *pruf* is a finite list of formulae, each of which is either i) an axiom, or ii) a formula $\mathcal{A}$, such that both "$\mathcal{B}$" and "$\mathcal{B} \Rightarrow \mathcal{A}$", for some formula $\mathcal{B}$, appear earlier in that list. A formula is a *thurem* if it is the last formula of some pruf.

Note that the prufs and thurems are both merely meaningless strings of symbols constructed in a certain way. They are not to be confused with the proofs the theorems of (informal) mathematics (which we think of as saying that "something is true"). You must now convince yourself that every argument you would accept as a proof in ordinary (informal) mathematics can be translated into a pruf, as defined above; and that every assertion you would accept as a theorem can be translated into a thurem. This is comparatively easy, once you have accepted that mathematical assertions can be translated into formulae, for the prufs and their thurems are structured just like the proofs and theorems of ordinary mathematics.

The above, then, outlines a scheme for formalizing mathematics.

Now let there be given some axiom system. Then there exists a Turing machine that will decide whether a list of formulae is a pruf or not (since that machine can check whether that list satisfies the conditions for a pruf). Hence, there is a machine that, given any integer S as input, will write out a thurem; and is such that every thurem is included in this list. [The machine simply tries lists of strings over $\tilde{\mathcal{C}}$ one at a time, checking for, and then reporting, those that are prufs.] That is, we can "mechanically generate all thurems". On the other hand, there is no obvious way to check, mechanically, whether a given formula is a thurem, for, although we can certainly write a Turing machine that looks for prufs of that formula, we have no way to determine whether or not that machine will halt.

The Gödel incompleteness theorem states that, for every such axiom system, one of two things is true. First, the system could be *inconsistent*. This means that there is some formula $\mathcal{A}$ such that both "$\mathcal{A}$" and "$\neg\mathcal{A}$" are thurems. Whenever this occurs, then (at least, for every reasonable axiom system) every formula becomes a thurem. Second, the system could be *incomplete*. This means that there is some formula $\mathcal{A}$ that is closed (i.e., is such that every free variable is subject to a "$\forall$"), and is such that neither "$\mathcal{A}$" nor "$\neg\mathcal{A}$" is a thurem. In informal terms, there is an assertion that is neither provable nor disprovable via the axioms. When this occurs, we could, of course, always add one of these to get a new axiom system — but then the incompleteness theorem again guarantees inconsistency or incompleteness of

that new system. The proof of the incompleteness theorem is like the proof that the halting problem is not computable. The crucial step is that the statements "there exists a pruf of $\mathcal{A}$" and "there does not exist a pruf of $\mathcal{A}$" can, using the set ω of integers, be reflected as formulae in the formal system (just as the crucial step in the halting problem is that Turing machines can query Turing machines).

As one example, let us take as our axioms a standard system, but without the axiom of choice (the formula that represents the assertion that, for every set x of disjoint sets, there exists a set having exactly one element in common with each element of x). Then it is known that neither the axiom of choice, nor its negation, is a thurem of that system. We may add the axiom of choice (or its negation!) as a new axiom.

9 Difficulty Functions

So far, we have been interested largely in which problems can be computed and which cannot. We now turn to a somewhat different set of issues, involving what resources are required for the computation process. These "resources" can be of several types, e.g., of memory space, of program length, or of time. We shall be interested in the last of these, for the benefit of utilizing quantum mechanics during the computation process appears to lie in the time required for that computation. It is entirely possible that there might be other benefits.

Let us begin with a simple example. Consider a (regular) Turing machine T, which computes some problem, π. Then for any string S, T, when run with S as the initial string, will eventually halt. Denote by $f(S)$ the total number of steps the machine T will execute before halting — a measure of the "time" required for the computation. We call this f the *step-difficulty function* of T. This function f clearly depends on the problem π itself; but may also depend on the particular algorithm we implemented (via T) in the computation of π. Note that every step-difficulty function satisfies $f(S) \geq 1$ for every string S.

With this example in mind, we now introduce the following definition. Fix a character set, $\mathcal{C}$. By a *difficulty function* over $\mathcal{C}$, we mean a function, $\mathcal{S} \stackrel{f}{\rightarrow} R$, from the $\mathcal{C}$-strings to the reals, which is positive and bounded away from zero, i.e., which, for some number $b > 0$, satisfies $f(S) \geq b$ for

every string S. Think of the number b as the time required to boot the computer: We do not wish to address the possibility that, for a couple of very simple input strings, the computer might be able to provide an answer in "zero time", or in an arbitrarily small time. The step-difficulty function of a Turing machine that solves a problem is, of course, just one example of a difficulty function. [Note that, while the step-difficulty functions are all integer-valued, we allow (for later convenience) our difficulty functions to be real-valued.]

While the above is of course merely a definition within mathematics, it is our intention to apply it to certain computations — both Turing and otherwise. In light of this intended application, we realize that this definition has an unfortunate feature: The difficulty functions provide too much detail. For example, it might be argued that a Turing machine should be allotted less time for a step in which the character under the head remains unchanged than for a step in which the machine has to print a whole new character. Or, we might purchase for our Turing machine a new chip, which runs twice as quickly as the old (but, say, takes longer to boot). These changes in the computing set-up would, arguably, require a different choice of difficulty function. But, while such technological improvements can certainly be important, they are not the subject of interest here. We, rather, are concerned with issues such as comparing, with respect to their difficulty, several problems, or several algorithms for computing the same problem. These ideas motivate the following definition: Given two difficulty functions, f and f', we write $f \sim f'$ provided that, for some number $a > 0$, $f(S) \leq af'(S)$ and $f'(S) \leq af(S)$ for all strings S. We note that this is indeed an equivalence relation on difficulty functions. It is the equivalence classes that reflect the sense of difficulty that we are concerned with here; and we shall always be interested in difficulty functions *only* up to this equivalence.

> **Exercise.** i) Fix a Turing machine, with step-difficulty function f, that solves a problem. Let f' be the difficulty function that results if the charge is only half a unit for a Turing-step that leaves the character on the tape unchanged, but still a full unit for a Turing-step that prints a new character. Prove that $f \sim f'$. A similar result holds for new allocations of units depending on the machine internal state, on whether the head is to be moved to the left or right, etc. ii) Prove that, for a any positive number, $f \sim af$

and $f \sim f + a$. iii) Prove that, for $a < \text{glb}(f)$, $f \sim f - a$ (where "glb" denotes the greatest lower bound). iv) Prove that, if f and f' are equal for all but a finite number of strings S, then $f \sim f'$. v) Characterize the functions h with the following property: Whenever $f \sim f'$, then $h(f) \sim h(f')$. vi) Let Turing machines T and T' compute problems π and π', respectively. Then we have seen how to build from these two a new machine, T'', that computes $\pi'' = \pi \circ \pi'$. Show that the corresponding difficulties (up to equivalence) are related by $f''(S) = f'(S) + f(\pi'(S))$.

These examples show, among other things, that the equivalence classes have some very desirable properties: The difficulty equivalence class does not depend on how units are allocated for various types of Turing steps, on how much time is required for booting, on the purchase of a better chip, on the act of learning how to treat a few S's very quickly.

We next introduce two notions that compare difficulty functions.

Let f and f' be two difficulty functions. We write $f \leq f'$ provided that, for some number $a > 0$, we have $f(S) \leq af'(S)$ for every string S. We note that: i) replacing f and f' by equivalent difficulty functions does not change this relationship; ii) both $f \leq f'$ and $f' \leq f$ hold if and only if $f \sim f'$; iii) $f \leq f' \leq f''$ implies $f \leq f''$; and iv) for f bounded above, we have $f \leq f'$ for every f'. That is, "$\leq$" has the properties one would associate with "less than or equal to". But note that, given two difficulty functions f and f', it is not necessarily the case that either $f \leq f'$ or $f' \leq f$. For example, on the positive integers, let $f(n) = \sqrt{n}$ and $f'(n) = 1 + n \sin^2(n/20)$.

There is, in addition to "$\leq$", a second type of inequality on difficulty functions. For f and f' two difficulty functions, we write $f \ll f'$ provided that, for every number $a > 0$, we have $f(S) \leq af'(S)$ for all but at most a finite number of strings S. We note that: i) replacing f and f' by equivalent difficulty functions does not change this relationship; ii) $f \ll f'$ and $f' \ll f$ cannot both hold; iii) $f \ll f' \ll f''$ implies $f \ll f''$; iv) $f \ll f'$ implies $f \leq f'$; and iv) either of $f \leq f' \ll f''$ or $f \ll f' \leq f''$ implies $f \ll f''$. Again, these are precisely the properties suggested by the notation. Since these special meanings of "$\leq$" and "$\ll$" relate only *functions*, there will be no confusion with the usual meanings of these symbols, which relate only *numbers*.

We think of $f \leq f'$, with $f \not\sim f'$, as meaning that "on every string, f

reflects no more difficulty than does f'; and there is an infinite number of strings on which f reflects strictly less difficulty". We think of $f \ll f'$ as meaning that "f reflects less difficulty than f' on every string". The following example will illustrate these ideas

> **Example.** Consider the palindrome problem of Sect. 5. Denote by f the step-difficulty function (counting steps) for the Turing machine T described in that Section. Set $L = \text{length}(S) + 1$, another difficulty function on $\mathcal{S}$. [The "+1" in this formula merely allows us to avoid treating the empty string separately.] Then we have $L \leq f \leq L^2$. The first follows because T must in any case traverse the entire string S (in order to examine the last character), and that traversal already requires L steps. The second follows because in the worst case, when S actually is a palindrome, T must go back and forth across the string (or a substantial portion thereof) a total of L times, together with a few extra steps at the ends. Note that these relations are not approximations: They hold *exactly*. Although $L \ll L^2$, we have neither $L \ll f$ nor $f \ll L^2$. Here is another Turing machine, $\tilde{T}$ for computing this problem. Machine $\tilde{T}$ works the same as T, except that, on the first pass, it makes an extra check to see if the string S is of the form "$aaa\cdots a$". If it finds that form, then $\tilde{T}$ immediately returns to the beginning and reports "yes". Denote by $\tilde{f}$ the step-difficulty function of $\tilde{T}$. Then, for every string S that is not all "a"'s, $\tilde{f}$ requires more steps than f (since T doesn't have to carry out those extra checks that $\tilde{T}$ does), but $\tilde{f}(S)$ and $f(S)$ differ at most by some numerical multiple of L (since this checking for "a"'s requires just a few extra steps for each character in S). However, for a string S that *is* all "a"'s, $f(S)$ is the order of L^2 (since T will have to go through the laborious process of checking for palindrome-ness), while $\tilde{f}(S)$ is the order of L (since $\tilde{T}$ will recognize this special form on the first pass). Note that there is an infinite number of such strings. It follows from all this that $\tilde{f} \leq f$, but neither $\tilde{f} \ll f$ nor $\tilde{f} \sim f$. We thus think of the computation represented by machine $\tilde{T}$ as "definitely (but only slightly) more efficient" than that of T. It seems plausible, intuitively, that: Given any Turing machine T' (step-difficulty f')

that computes this problem, there exists a Turing machine T'' (step-difficulty f'') that also computes this problem, such that $f'' \leq f'$ and $f'' \not\sim f'$. This means that, no matter how efficient you feel your present Turing machine is, there always exists one that is a little more efficient. It would be interesting to find a proof. On the other hand, it is known ([**?**]) that there exists no Turing machine T' (step-difficulty f') that computes the palindrome problem and has $f' \ll L^2$. In other words, there is no way to compute the palindrome problem with efficiency substantially greater than L^2.

This example illustrates the idea that this equivalence relation and these inequality-relations on difficulty functions are the "right" notions: They allow us to express, in a simple way, what we want to say; and they don't draw us into a discussion of what we don't want to say.

One could imagine inventing other, inequality-like, relations on difficulty functions. For example, one could compare averages of the values of the functions over certain strings; or consider the relative frequencies of the S's for which $f(S) \leq f'(S)$ or $f'(S) \leq f(S)$ occur. But these relations tend not to be very interesting, probably because they typically require some choice of an ordering for the strings, or they are are too sensitive to relatively benign relabelings of the strings.

10 Difficult Problems — Best Algorithms

In this section, we discuss two results of Blum [**?**]. Both of these results are insensitive to the particular difficulty-measure — or even language — employed; and both are proved by diagonal arguments. For ease of exposition, we shall discuss both results for the Turing case (i.e., the "machines" will be Turing machines, and the difficulty measures will be step-difficulty). But it should be noted that these restrictions are not necessary.

A "difficult" problem is, intuitively, one that requires many steps for its computation. It is easy to think of problems that appear, offhand, to be quite difficult in this sense, e.g., that which sends any integer S to the integer that is the third-to-last digit of the $(10^{S!})!$-th prime. But it is hard to be certain that this problem really is as difficult as it appears: There might,

for example, be some marvelous theorem that asserts that this particular problem π merely returns "7" when S is even; and "1" when S is odd. If this, or something like it, should turn out to be the case, then this problem π would turn out to be easy to compute.

Can we give an example of a problem π that is computable, and is such that we can *guarantee* that any Turing machine that computes it has step-difficulty, say, $\geq (10^{S!})!$) ? The answer is yes, but for a silly reason. Let π be the problem that, applied to positive integer S, returns the string $a \cdots a$, where the total number of a's is $(10^{S!})!$. Then certainly the step-difficulty f of any Turing machine that computes this π satisfies $(10^{S!})! \leq f$, since it takes this many steps for the Turing machine merely to print out (never mind compute) the answer. This isn't exactly what we had in mind. So, to avoid this sort of foolishness, we introduce the following definition. A problem π will be said to be *bounded* if the lengths of the strings $\pi(S)$ as S ranges through all input strings, are bounded above.

So, are there very difficult — perhaps even "arbitrarily difficult" — bounded problems? We formulate this question precisely as follows:

Assertion. Let $\tilde{f}$ be any difficulty function. Then there exists a bounded, computable problem π with the following property: Every Turing machine T (step-difficulty f) that computes π has $f \geq \tilde{f}$.

This assertion states, in other words, that you tell me how hard ($\tilde{f}$) you want the problem to be, and I'll find a problem (π) such that every method of computing it (T) is at least $\tilde{f}$-difficult (i.e., has $f \geq \tilde{f}$). Unfortunately, this assertion turns out to be false.

Here is a counterexample. First note that, for any Turing machine T that computes a problem, the step-difficulty function, f, of that machine is computable. Indeed, consider the Turing machine T' that, on any string S, merely simulates the action of T on S, counting the number of steps T runs before halting, and reporting that number. This T' computes f. Now let $f_1, f_2, \cdots$ be a list of all computable, integer-valued difficulty functions (noting that the collection of such functions is countable, since the collection of all Turing machines is already countable); and let $S_1, S_2, \cdots$ be a list of all strings. Now define a new function, $\tilde{f}$, on strings by

$$\tilde{f}(S_n) = n \times max[f_1(S_n), f_2(S_n), \cdots, f_n(S_n)]. \tag{3}$$

[In other words, the value of $\tilde{f}$ on string S_n is n times the largest of the values taken by the first n of our computable functions f_i, acting on that S_n.] This $\tilde{f}$ is our counterexample. To see this, fix any positive integer m. Then, for any $n \geq m$, we have $\tilde{f}(S_n) \geq n f_m(S_n)$ (for, since $n \geq m$, f_m is included in the functions maxed-over in (**??**)). But this last inequality (for all $n \geq m$), and the fact that there is only a finite number of $n < m$, imply $f_m \ll \tilde{f}$. But the f_m exhaust step-difficulties of Turing machines that compute problems, and so there can be no such difficulty function f satisfying $f \geq \tilde{f}$.

The idea of this example is to so construct $\tilde{f}$ that it "grows very quickly as the string S gets larger — so quickly that no computable, integer-valued function (and therefore certainly no Turing step-difficulty function) can keep up with it". This growth is very fast indeed, for we can think of some pretty fast-growing computable f's, e.g., (for S an integer) $f(S) = 2$ to the power of 2 to the power of 2 . . . S times. Well, that particular f (being, as it is, computable) is child's play in the hands of the *really* fast-growing $\tilde{f}$ of the theorem. This situation may seem paradoxical at first sight: How can $\tilde{f}$ grow more quickly than any computable function, when Eqn. (**??**) appears to be a computation of $\tilde{f}$? But closer inspection reveals that we do not actually "compute" $\tilde{f}$ above, for we cannot Turing-construct a list of Turing machines, $T_1, T_2, \cdots$, that compute the original list $f_1, f_2, \cdots$ (without, that is, computing the halting problem). Yet, although we cannot Turing-construct this sequence, it certainly does exists, for the set of all Turing machines is already countable, and so therefore is the set of all computable integer-valued difficulty functions.

So, to summarize, it is possible to invent absurd levels of difficulty (such as that described by the $\tilde{f}$ of the example above): There exist no computable problems that are *that* difficult. But what about more reasonable levels of difficulty? One might think of demanding that $\tilde{f}$ be computable, for, by the remarks above, this condition would prevent $\tilde{f}$ from growing too fast. But this demand must be implemented with some care. First, difficulty functions are real-valued, and so "computable" does not make sense for them. [Indeed, most real numbers are not even computable.] But note that every difficulty function $\hat{f}$ is equivalent to an integer-valued one (namely the function whose value, on each string S, is the smallest integer exceeding $\hat{f}(S)$); and, for integer-valued difficulty functions, "computable" certainly does make sense. This suggests that we demand, of the function $\tilde{f}$ of the assertion, that it be integer-valued and computable. It turns out that, with this additional

condition, our assertion is true:

Theorem. (Blum) Let $\tilde{f}$ be any computable, integer-valued, difficulty function. Then there exists a bounded, computable problem π with the following property: Every Turing machine T (step-difficulty f) that computes π satisfies $f \geq \tilde{f}$.

Proof: Let $T_1, T_2, \cdots$ be a list of all Turing machines (over the given character set), and $S_1, S_2, \cdots$ a list of all strings. Now fix any positive integer n, and consider the following prescription (ignoring for the moment the words in braces):

> **Prescription**(n)**:** Attempt to run each of the first n machines, $T_1, \cdots, T_n$, in this order, on the initial string S_n, for a total of $\tilde{f}(S_n)$ steps each. If none of the {uncanceled} machines, so run, halts before reaching $\tilde{f}(S_n)$ steps, set $\pi(S_n) = \emptyset$. Otherwise, denote by T_i the first {uncanceled} machine in this list that *does* halt before reaching $\tilde{f}(S_n)$ steps[4]. Then set $\pi(S_n)$ equal to a string other than T_i's output: Set $\pi(S_n) =$ "a" if T_i, on S_n, halted with output string $\emptyset$; and $\pi(S_n) = \emptyset$ otherwise. {Finally, cancel that T_i.}

This prescription, carried out for all values of n, defines a problem π (since it prescribes what string, $\pi(S)$, is to be assigned to each string S). We note that this π, so defined, is bounded (since its only possible output strings are $\emptyset$ and "a"). Furthermore, this π is computable. This follows because we can build a Turing machine that i) produces the sequence $T_1, T_2, \cdots$ of (*all*) Turing machines and the sequence $S_1, S_2, \cdots$ of all strings; ii) simulates the running of the first n machines, as in the prescription; iii) finds the first {uncanceled} machine that fails to run for at least $\tilde{f}(S_n)$ steps (here, using the fact that $\tilde{f}$ is computable!); and iv) sets $\pi(S_n)$ accordingly.

We now reinstate the braces. We carry out the prescription above, in turn, for successive values of n: $1, 2, \cdots$. Each time this prescription (for some n-value) is carried out, that machine T_i (if any) used to set $\pi(S_n)$ is

[4]Note that T_i is the first machine in this ordered listing of machines that halts at all before the $\tilde{f}(S_n)$ steps; and *not* that machine, among these n machines, that halts in the fewest steps

now "canceled", i.e., excluded from consideration in subsequent (i.e., larger-n) applications of the prescription. Thus, the new construction is identical to the old, except that, because of this cancellation, the list of Turing machines included at each stage may be smaller than it was before. But in any case the result is again some bounded, computable problem, π (different from the old π, with which we are no longer concerned).

This π is the problem whose existence is guaranteed by the theorem. To see that it has the required property, consider any Turing machine T (step-difficulty function f) that computes π. Then this T must appear somewhere in our list of machines: Say, $T = T_7$. Consider the machines $T_1, \cdots, T_6$. Let n_o be an integer such that every one of these six machines either was canceled already by the time n reached n_o; or never will be canceled for any n. [Such an n_o exists: Indeed, each of the machines $T_1, \cdots, T_6$ either i) is at some point (i.e., for some specific n-value) canceled, or ii) is never canceled. Let n_o be the largest of the specific n-values that occur in i).] Now fix $n > n_o$, and apply the prescription above to determine $\pi(S_n)$. Which, if any, machine is canceled during this application of the prescription? It could not be any of $T_1, \cdots, T_6$ (by definition of n_o). Therefore, T_7 is on the bubble: It will not be saved by cancellation of any of the machines before it in the list, and so will be canceled if it halts before completing all $\tilde{f}(S_n)$ steps. But T_7 cannot be the one canceled either, for, by definition of $\pi(S_n)$, cancellation implies that T_7 on S_n differs from $\pi(S_n)$, while T_7 was assumed to compute π. We conclude from all this that T_7, on S_n, must run for at least $\tilde{f}(S_n)$ steps without halting. That is, we conclude that $f(S_n) \geq \tilde{f}(S_n)$. Since this holds for all $n > n_o$ (i.e., for all but at most a finite number of n), we conclude that $f \geq \tilde{f}$. \

This is quite a proof. For each n, we stage a contest between the first n Turing machines, applying each to S_n and seeing who can go at least $\tilde{f}(S_n)$ steps without halting. We find the first machine that fails, arrange for π to be different from what *that* machine computes, remove that machine from further competition, and then repeat the contest for the next n. Since $\tilde{f}$ generally increases, the successive contests will generally get harder and harder. In this way, π avoids the losers (the machines that halt early), and thus emerges as a problem that can only be computed by a consistent winner — a machine with step-difficulty satisfying the condition of the theorem. Note that the n_o in the proof is not computable. Note also that computability of $\tilde{f}$ is used at a critical place: To get computability of π.

Exercise. Show that the theorem above continues to hold if the last formula in its statement is replaced by $f \gg \tilde{f}$. Does there exist a Turing machine that accepts as input the Turing machine that computes $\tilde{f}$, and returns a Turing machine that computes a problem π whose existence is guaranteed by the theorem?

So, there are some pretty hard problems out there. We now turn to a related issue. It would be of great interest to define, for any given problem, a difficulty intrinsic to a *problem itself* (rather than to whatever method is currently being used to compute that problem). A possible line for introducing such a notion would be to let the "intrinsic difficulty" of a problem mean the minimum step-difficulty function of Turing machines that compute that problem. But, in order to implement such an idea, we would need some result to the effect that this minimum is actually achieved. One result that would certainly do the trick is the following:

Conjecture. Let π be any computable problem. Then there exists a Turing machine T (step-difficulty f) that computes this problem, with the following property: Given any other Turing machine T' (step-difficulty f') that computes this problem, we have $f' \geq f$.

Then we would take the f of the conjecture as a measure of the intrinsic difficulty of the problem π. Unfortunately, this conjecture is false. Indeed, even for the case of the palindrome problem we have observed that, for any Turing machine T (step-difficulty f) we can think of offhand to compute this problem, there exists another, T' (step-difficulty f') with $f' \leq f$ and not $f' \sim f$. [Embarrassingly enough, we don't understand even this simple little problem well enough to generate from it an actual counterexample to the conjecture above!] Here, however, is a possible alternative conjecture — weaker than the one above, but perhaps retaining enough strength to salvage some sort of notion of intrinsic problem-difficulty.

Conjecture. Let π be any computable problem. Then there exists a Turing machine T (step-difficulty f) that computes this problem, with the following property: There is no Turing machine T' (step-difficulty f') that computes this problem, such that $f' \ll f$.

This conjecture, for example, is true for the palindrome problem. Indeed, we gave a Turing machine that computes this problem with a difficulty function f with $f \leq L^2$, but not $f \ll L^2$; and it is known ([**?**]) that there is no Turing machine that computes this problem with difficulty $\ll L^2$.

Thus, this last conjecture looks promising. But, much as we might wish it to be otherwise, this conjecture is also false. Indeed, we have

Theorem. (Blum) There exists a computable problem π with the following property: Given any Turing machine T (step-difficulty f) that computes π, there is another Turing machine T' (step-difficulty f') that also computes π, such that $f' \ll f$.

Thus, according to this theorem, for this particular problem π, no matter how much effort you put into finding an efficient machine of computing π, there always exists a much more efficient machine waiting in the wings. You can, if you wish, submit that new, more efficient machine to the theorem, and it will then go ahead and guarantee the existence of a still more efficient machine, and so on. Thus, for the problem π of the theorem, there is an infinite succession ever more efficient Turing machines that compute it. There would seem to be no hope of defining an "intrinsic difficulty" for this problem, at least.

We shall merely sketch the proof of the theorem. First, let h be the integer-valued function on nonnegative integers defined by: $h(0) = 1$, and, for $n > 0$, $h(n) = 2^{(h(0)+\cdots+h(n-1))}$. This function is rather rapidly-growing: $h(1) = 2$; $h(2) = 8$; $h(3) = 2048$, $h(4)$ would take about ten lines to write out; and $h(5)$ could not be written on all the paper ever manufactured on this planet. We next construct the problem π of the theorem, as follows. This construction is identical with the construction of the problem π in the proof of the previous theorem, with just one small change. In carrying out the prescription, for some n-value, instead of running each of the machines $T_1, T_2, \cdots, T_n$ for the same number, $\tilde{f}(S_n)$, of steps, we now run the i-th machine in this list for $h(n-i)$ steps. Thus (since h is rapidly growing), the early machines in the list are run for vastly more steps (to see if they halt) than are the later machines in the list. In any case, the result, after this one change, is a certain computable problem π (different, of course, from the π of the previous theorem). This π, believe it or not, has the property required in the theorem.

To see this, let Turing machine T (step-difficulty f) compute π. Then this T must be one of the T_i in our list, say $T = T_7$. By construction, using the same argument as in the previous proof, it follows that $f(S_n) \geq h(n-7)$ for every n. We now introduce a new problem, π'. We set $\pi'(S_1) = \emptyset, \cdots, \pi'(S_7) = \emptyset$. For $n > 7$, we define $\pi'(S_n)$ by exactly the same prescription that defined π above, except that we use for our list of machines, not the $T_1, \cdots, T_n$ as was used above, but rather just $T_8, T_9, \cdots, T_n$. This π' is of course also computable.

We next note that π' and π are actually equal on all but at most a finite number of strings. This follows because for sufficiently large n, say, $n \geq n_o$, each of $T_1, \cdots, T_6$ that ever will be canceled in the computation of π has already been canceled (while T_7, of course, will never be canceled). Once no more cancellation of these seven machines is possible, then π' and π are left to examine precisely the same machines at each step, namely, $T_8, \cdots T_n$, and so these two will end up with the same values.

We next introduce a Turing machine T' that computes π in the following manner. For $n \leq n_o$, T' simply simulates T, in this way finding out what $\pi(S_n)$ is, and returns that string. On the other hand, for $n > n_o$, T' computes π' in the manner described above (i.e., using, in the prescription at each stage, only machines $T_8, \cdots, T_n$). Denote by f' the step-difficulty of this T'.

Finally, we claim that $f' \ll f$. It suffices to compare these two difficulty functions on S_n with $n > n_o$ (since these S_n include all but a finite number of strings). Fix $n > n_o$. Then, in order to compute $\pi'(S_n)$ $(= \pi(S_n))$, machine T' must simulate Turing machine T_8 (on S_n) for $h(n-8)$ steps, machine T_9 for $h(n-9)$ steps, and so on up to machine T_n for $h(0)$ steps. Thus, T' must run a total of not more than $h(0) + h(1) + \cdots + h(n-8) = \log_2(h(n-7))$ steps, where the last equality follows from the construction of h. Thus, we conclude that for $n > n_o$, $2^{f'(S_n)} \leq h(n-7) \leq f(S_n)$, where the last step is the bound on $f(S_n)$ found earlier. The result follows.

The first thing to notice about this argument is that it contains a flaw: Right at the end, we are comparing the number of steps that T *actually executes* with the number that T' must *simulate*. Simulating looks like a lot more work that merely executing. But for reasonable difficulty-measures in reasonable languages (although not for step-difficulty in Turing) a machine can be simulated in the same number of steps (up to equivalence) as it can be run. In these cases, which include all those of serious interest, the argument is complete. But for the Turing case, a further, somewhat complicated,

workaround is necessary, which we shall not discuss.

Actually, we prove more than is stated in the theorem, namely that $2^{f'} \leq f$. In fact, one can obtain a similar result for other, specific, choices of an inequality relating f and f', by simply changing the choice of the function h. It is interesting to note that, although the theorem guarantees the *existence* of T', it does not tell us how to compute it. The crucial non-computable step is that in which n_o is found. In fact, there exists no Turing machine that, with input a Turing machine T that computes the π of the theorem, returns a Turing machine T' the existence of which is guaranteed by the theorem.

So, to summarize, the prospects for assigning to each problem an "intrinsic difficulty", in some reasonable way, look pretty dismal. It may be possible to do better by some appropriate restriction on the class of problems considered. Or, there may be some way to take the greatest lower bound of the difficulty functions for machines that compute the problem, even though that lower bound is itself not realized by any machine.

11 A Language for Efficiency

Clearly, Turing machines are highly inefficient. The key problem is that storing scratch work on a single long tape requires that the machine plod, again and again, over the same portion of tape, looking for one little piece of data after another. In the case of the palindrome problem, for example, no Turing machine can compute this problem in step-difficulty $\ll L(S)^2$; and yet we might expect a "normal" computer to require only $L(S)$ steps. Thus, Turing step-difficulty functions tell us too much about Turing language and too little about the subject of real interest: the "intrinsic difficulty" of the problem or algorithm. It is time to upgrade.

We might do so, e.g., to Fortran. We would assign, in some "reasonable" way, a number of steps to each Fortran command; and thereby arrive at a Fortran-difficulty function for each Fortran-computed problem. We could, of course, do the same for C language, etc. While these new difficulty functions would certainly be more realistic than Turing step-difficulty, there remains the danger that they, too, would manifest excessive language-dependence. But it seems, intuitively, that such dependence may be small, or — if things are set up carefully — even absent. One might imagine, for example, that we could write a C-emulator in Fortran that is difficulty-function preserving.

This situation with respect to difficulty, then, is very like that we faced earlier with respect to computability: There appears to be a universal notion lurking in the background, but that notion finds expression through many languages. We want to distill out the notion itself. The answer, in the case of computability, was Turing machines. We find the simplest language that is still rich enough to encompass our idea of computability, and then *define* computability in terms of that language. We would now like to do the same thing for difficulty. That is, we would like to invent a language, with an associated difficulty function, that is as simple as possible, but not so simple that it generates unnecessary inefficiencies. In short, we want to find a language that is to difficulty as Turing language is to computability. It will turn out, unfortunately, that our innate sense of what is the "correct" difficulty function is somewhat less firm that that of what is "computable". But, in any case, we propose, below, a language that seems to capture a more or less reasonable notion of "difficulty". There may very well be better proposals.

Fix a character set $\mathcal{C}$. For S any string over $\mathcal{C}$, we write $L(S)$ for the number of characters in the string S plus one [The "plus one" is so we don't have to treat $S = \emptyset$ as an exception.]

Let there be created an infinite number of storage locations, each labeled by some string over $\mathcal{C}$; and each capable of holding an arbitrary string over $\mathcal{C}$. Thus, we impose no upper bound on the number of storage locations being utilized, nor on the lengths of the strings in the various locations (although each location, at any one moment, contains merely a string, i.e., a finite sequence of characters; and it will turn out that only a finite number of storage locations are in play at any one moment). We write $C(S)$ for the string in the location labeled S. The idea here is that in this way we create an ample amount of highly accessible storage space.

In the present language there will be commands, each of which directs that a certain action (mostly involving what is stored in certain locations) be taken. There is a total of five classes of commands in this language: two for input/output; two for manipulating strings; and one for branching. Listed below are these five classes of commands (with, for each, a brief explanation of what is to be done; and, in braces, a number representing the "difficulty" of the command, which we shall discuss shortly).

A command results if, in any of the five items below, "S" is

replaced by any explicit string, "x" by any explicit character, and "n" by any (positive or negative) explicit integer:

1. INPUT TO $C(S)$: allows the user to enter any string, which is then placed in location S. {L(whatever string is entered)}]

2. OUTPUT FROM $C(S)$: allows the user to retrieve the string stored in location S. {L(C(S))}

3. APPEND x TO $C(C(S))$: replace whatever string is stored in location $C(S)$ with that same string, but with character x appended on the right. {L(C(S))}]

4. DELETE LAST OF $C(C(S))$: replace whatever string is stored in location $C(S)$ with the string that results from deleting its rightmost character (if any). If $C(C(S)) = \emptyset$, do nothing. {L(C(S))}

5. IF (LAST $C(C(S)) == x$) SKIP n LINES: if the last character (if any) of the string in location $C(S)$ is "x", then skip forward n program lines (if n is positive), backward $|n|$ lines (if negative). If $C(S) = \emptyset$, or if the last character of $C(S)$ is other than x, or if the line to be skipped to is an INPUT command, or if there are insufficient lines in the program to carry out the indicated skip, do nothing. {L(C(S))}

By "explicit" above, we mean that the character x or the string S or the integer n must actually be written out, within the command, as some specific character or string: It cannot be indicated only implicitly, e.g., as whatever happens to be stored in some location.

A *program* is a finite ordered list of commands, with the following property: The program contains exactly one INPUT command, and it is the first command of the list; and exactly one OUTPUT command, and it is the last command of the list. Here is an example of a program

INPUT $C(abc)$
APPEND d TO $C(C(yzr574))$
IF (LAST $C(C(m)) == a$) SKIP -1 LINES
OUTPUT $C(yes)$

To run a program, place $\emptyset$ in every storage location, begin at the first program line (INPUT), and enter any string. The machine then carries out the

instruction of each command in turn, then moving on to the next command in the list (except for the case of command 5 (IF), for which the next command to be executed is the one indicated above). If and when the machine reaches the last command (OUTPUT) of the list, the machine halts, allowing the user to read the output string.

Any program, run on any input string, either halts or does not halt. If it halts for every input string, then that program *computes* some problem π, where $\pi(S)$ is the output string when string S is entered at INPUT. The program above, for example, indeed computes a problem, namely that with $\pi(S) = \emptyset$ for every string S. Note that we have so structured the commands that the program cannot "hang" within a single command: As long as the command follows the grammatical rules above, then — no matter how pointless that command might be — the machine will always do *something* (or maybe nothing, as the case may be) and move on. Failure to halt can only occur by continuing to execute command after command, indefinitely, as in the following example:

INPUT $C(x)$
APPEND d TO $C(C(x))$
IF (LAST $C(C(x)) == d$) SKIP -1 LINES
OUTPUT $C(x)$

We could have modified the way the IF command works, in the following manner. We could, first, require that each command in the program be labeled by a unique string. Then, we could revise the command IF so that it directs, not that some number of program lines be skipped, but rather that there be executed next that command with some explicit string-label. Clearly, this modification adds nothing new.

The numbers in braces, accompanying each of the five commands above, give the number of "steps" we deem the computer to require to execute that command. We call this number the *difficulty* of the command; and, for a program that, acting on a certain string, halts, we call the total number of steps executed the *difficulty* of that run of the program; and, for a program that computes a problem, we call the total number of steps executed before halting (a function now of the input string) the *difficulty function* of the program. As always, we are interested in difficulty functions only up to equivalence. There follows a discussion of the difficulties assigned, above, to

the five classes of commands.

If, in response, to an INPUT command, there is entered a string of, say, 13 characters, then the execution of that command requires, as dictated above, 14 steps. This surcharge for entering long strings turns out to be very convenient (e.g., already in the following paragraph).

The number of steps assigned to the OUTPUT command is L(the string returned). This formula was chosen merely for aesthetics: Even changing it, e.g., to "1" would result in equivalent difficulty functions. To see this, first note that, for any program on any string that runs up to the OUTPUT command, the total difficulty up to that point will be greater than or equal to the length of the longest string stored. [This follows since each command adds at least as much to the cumulative difficulty as it adds to the length of the longest string.] Thus, changing the difficulty for the OUTPUT command to "1" would, at most, reduce the total difficulty function by a factor of two. But such a reduction results in an equivalent difficulty function.

For the APPEND command, we append a character to the string in the location given by the string in the location S. We have to look up location S, to find $C(S)$, and then look up location $C(S)$ to find the string to be appended. Think of the difficulty, $L(C(S))$, of this command as a "lookup charge". Why is not the formula instead $L(C(S)) + L(S)$, i.e., why don't we also have a charge for "looking up" S? The reason is that this change results in an equivalent difficulty function. Indeed, in any given program there will be a finite number of APPEND commands, and so a finite number of explicit strings S in those commands. So, there will be a longest such string, say seven characters. Thus, a change in the difficulty of APPEND to $L(C(S)) + L(S)$ will add at most eight steps for this command, i.e., will increase the difficulty for this command by a factor of at most nine. As a result, the final difficulty function for this program will increase by at most a factor of nine. But such an increase results in an equivalent difficulty function. Note that the same argument does not apply to the term $L(C(S))$ in the difficulty of APPEND: This number (one more than the number of characters stored in location S) depends on what happens to be stored in S at the time, and so cannot be bounded a priori. Thus, this term may make a nontrivial contribution to the final difficulty function. Why not include a term $L(C(C(S)))$ in the difficulty function of the APPEND command? After all, we have to travel to the end of the string $C(C(S))$ to append the x, and there should be some travel allowance. This is a reasonable position, which might be worth pursuing.

But we have to make some decision here, and we have elected the viewpoint that there has been constructed some sort of pointer that allows us to find the end of the string easily. Similar remarks apply to the DELETE and IF commands.

In the IF command, why not include also a term $|n|$, the number of command-lines skipped? After all, skipping lines is hard work, and there should be some compensation. But, again, a given program has but a finite number of IF commands, each with an explicit n; and so a maximum value of $|n|$ for all such commands in the program. Therefore, such a change in the difficulty for the IF command would always result in an equivalent difficulty function.

We now have to deal with two matters. First, we need to show that the computable problems in this new language are precisely the computable problems (defined earlier, using Turing machines). And, second, we would like to argue that the difficulty functions generated by this language are "reasonable", i.e., that they correctly capture our intuitive sense of what the difficulty "should" be. We shall attempt to resolve both of these matters in one sweep, by generating a list of illustrative subroutines — i.e., of short program-fragments. On the one hand, these subroutines will show the richness of what can be computed in this language. On the other hand, the difficulties of these subroutines, computed from the command-difficulties above, will illustrate the typical difficulty functions this language generates. In these subroutines, $S, S', \cdots$ stand for any explicit strings, x for any explicit character, and n for any explicit integer.

L. APPEND x TO $C(S)$. $\{1\}$
2. DELETE LAST OF $C(S)$. $\{1\}$
3. IF (LAST $C(S) == x$) SKIP n LINES. $\{1\}$

For subroutine 1, let, say, $S =$ "yzr". First, rewrite the program, if necessary, so that location "$\emptyset$", as well as some location, say "$h8$", are not used elsewhere in that program, so $C(\emptyset) = \emptyset$ and $C(h8) = \emptyset$. Then: APPEND y TO $C(C(h8))$; APPEND z TO $C(C(h8))$; APPEND r TO $C(C(h8))$; APPEND x TO $C(C(\emptyset))$; DELETE LAST OF $C(C(h8))$; DELETE LAST OF $C(C(h8))$; DELETE LAST OF $C(C(h8))$. The first three lines achieve $C(\emptyset) =$ "yzr"; the last three restore $C(\emptyset)$ to $\emptyset$. The total difficulty, for any one instance of this subroutine, is some fixed integer (in the example above, 10), depending only

on the explicit string S, and not on what is in the various memory locations at the time. But this subroutine can appear at most a finite number of times in any program, and so the actual difficulty contributed by this subroutine, each time it is run, is bounded above. So, we may assign this subroutine a difficulty 1, up to equivalence of difficulty functions. Similar remarks apply to subroutines 2 and 3.

4. SKIP n LINES. $\{1\}$
5. IF $(C(S) == \emptyset)$ SKIP n LINES. $\{1\}$
6. IF $(C(S) == C(S'))$ SKIP n LINES. $\{L(C(S)) + L(C(S'))\}$
7. SET $C(S) = \emptyset$. $\{\mathrm{L(C(S))}\}$

For subroutine 4, use APPEND a TO $C(\emptyset)$; IF (LAST $C(\emptyset) == a$) SKIP n LINES; and then place DELETE LAST OF $C(\emptyset)$ as the first command executed after the skip. For subroutine 5, use the commands IF (LAST $C(S) == x$) SKIP .. as x runs over all possible characters; arranging the skips so that we skip n lines if all the IF's fail, but merely proceed to the next line if any succeeds. The difficulty of this subroutine, 1, results from the fact that the total number of characters is fixed. For subroutine 7, use, repeatedly, DELETE LAST FROM $C(S)$), in conjunction with subroutine 5 (to test whether $C(S)$ is empty yet). Note that subroutine 7 has a variable difficulty: Its value depends on how many characters will have to be removed from $C(S)$.

For the subroutines below, we suppose that we begin with $C(S) = \emptyset$. [If this location were not empty, then it would be necessary to use subroutine 7 first, to achieve $C(S) = \emptyset$; and to adjust the difficulty appropriately.]

8. SET $C(S) = S'$. $\{1\}$
9. SET $C(S) = C(S')$. $\{L(C(S'))\}$

For subroutine 9, we first use IF (LAST $C(S') == x$) SKIP ..., skipping to the command APPEND x TO $C(h8)$ (where "$h8$" is some location with $C(h8) = \emptyset$). Continue to test in this way each possible candidate, x, for the last character of $C(S')$. Then DELETE LAST OF $C(S')$, test whether $C(S') == \emptyset$ (subroutine 5); and, if not, repeat. In this way, we place $C(S')$, with its characters in reverse order, into $C(h8)$. Now do this all again, placing $C(h8)$, in reverse order, into $C(S)$. It should be clear at this point that we can carry out complicated string-manipulations, e.g.: Place in $C(S)$ every

other character of $C(S')$, up to the first occurrence of "a", and with each "c" replaced by "$8k$", provided that $C(S'')$ contains at least 6 characters not including the combination "yzr"; otherwise ...

For the next three subroutines, we assume that the digits, $0, 1, \cdots, 9$, are included in the character set; that strings subject to arithmetic operations are already integers; and that, again, we begin with $C(S) = \emptyset$.

10. SET $C(S) = C(S') + C(S'')$. {L(C(S'))+L(C(S''))}
11. SET $C(S) = C(S') * C(S'')$. {L(C(S'))*L(C(S''))}
12. SET $C(S) = L(C(S'))$. {L(C(S'))}

For subroutine 10, for example, we first use IF (LAST $C(S') == x$) SKIP ... and IF (LAST $C(S'') == y$) SKIP ..., for the one hundred possible combinations of digits substituted for x and y; placing, for each combination, the appropriate digit in $C(h8)$, say, as well as a marker in $C(h9)$, which tells whether or not we are carrying the 1. Then DELETE LAST OF $C(S')$; DELETE LAST OF $C(S'')$, test whether either $C(S') = \emptyset$ or $C(S'') = \emptyset$, and repeat. We will end up with the sum, with digits in reverse order, in $C(h8)$. Now transcribe $C(h8)$ into $C(S)$, reversing the order of digits. For subroutine 11, use the usual pencil-and-paper multiplication (in the course of which each digit of $C(S')$ must be multiplied by each digit of $C(S'')$). Using similar techniques, we can write write subroutines for loops, e.g., WHILE and DO; and also for complicated branchings, such as IF ((... AND NOT ...) OR ...) CARRY OUT ...; ELSE CARRY OUT ... .

It should be clear by this point that a problem is computable in this language if and only if it is (Turing) computable. After all, we have in this language the ability to enter and recover strings (INPUT, OUTPUT), the ability to manipulate strings (APPEND, DELETE) freely, and the ability to branch (IF).

Note that all of the subroutines 4-12 were constructed solely from commands 1 and 2 together with subroutines 1-3. That is, we have not so far used commands 3-5, other than to construct subroutines 1-3. Why, then, did we not omit commands 3-5, making our basic commands consist instead of commands 1 and 2, together with subroutines 1-3? The role of the "$C(C(S))$" in commands 3-5 is to allow indexed arrays (which, as we shall see shortly, play a role in efficient programming). Here are three subroutines that use this feature in an essential way:

13. SET $C(S'1)$ = FIRST CHARACTER OF $C(S)$, $C(S'2)$ = SECOND CHARACTER OF $C(S)$, ETC. $\{L(C(S))\ \log(L(C(S)))\}$
14. SET $C(S) = C(C(S'))$. $\{L(C(S')) * L(C(C(S')))\}$
15. SET $C(S) = C(C(C(S')))$. $\{L(C(C(S')))*(L(C(S'))+L(C(C(C(S'))))))\}$

In subroutine 13, we are assuming that the character set contains the digits; and "$S'2$" means the string resulting from appending the character "2" to the string S', etc. Thus, this subroutine allows us to place the individual characters of the string $C(S)$ in separate locations. This makes those characters directly accessible (without having to go through all of $C(S)$ each time a character is needed). The factor "$\log(L(C(S))$" in the difficulty reflects the fact that the length of the locations ($S'n$, for $n = 1, 2, \cdots$) increases logarithmically as the length of $C(S)$. Note that the base for this logarithm is irrelevant, up to equivalence. Subroutine 15 shows that we can index arrays with indexed arrays. This subroutine is given by SET $C(e8k) = C(C(S'))$; SET $C(S) = C(C(e8k))$; and this construction yields the indicated difficulty.

> **Exercise.** Explain how to write a subroutine SKIP $C(S)$ LINES (which, say, does nothing if $C(S)$ is not an integer). What is its difficulty?

This concludes our treatment of the present language. We conclude this section with a few additional remarks.

First note that, for any program that computes a problem, the difficulty function is $\geq L(S)$. This follows, since INPUT already imposes a difficulty equal to the length of the string entered plus 1. Next, consider two programs, which compute problems π and π'. Then it is easy to write a program that computes problem $\pi \circ \pi'$: Simply juxtapose the two programs, and remove the two lines where the OUTPUT of one abuts the INPUT of the other (and, possibly, change a few explicit strings). The difficulty of the new program is the sum of the difficulties of the two components. It is easy to write short programs that change the grammar of inputs and outputs: Encoding "yes" and "no" in different ways, changing the number base, changing character set, using character-orderings in various ways, rejecting uninteresting inputs, inserting and removing separators, etc. These always have difficulty $L(S)$,

where S is the string entered. It follows from all these remarks, taken together, that the difficulty function (up to equivalence) is independent of the input-output grammar.

For f and f' difficulty functions, denote by $\text{glb}(f, f')$ the function whose value, for each string S, is the smaller of the values of $f(S)$ and $f'(S)$. Then $\text{glb}(f, f')$ is also a difficulty function, and, up to equivalence, depends only on the equivalence classes of f and f'. We have: $\text{glb}(f, f') \leq f$ and $\text{glb}(f, f') \leq f'$; and $\text{glb}(f, f') \sim f$ if and only if $f \leq f'$. Now let π be a problem, and let P (difficulty function f) and P' (difficulty function f') be programs that compute π. Then there exists a program, P'', that computes π, with difficulty function $\text{glb}(f, f')$. This P'' is constructed as follows. Program P'' first makes a copy of the initial string S, then simulates the running of P on S for ten steps; then the running of P' on the copy of S for ten steps; then continues the simulation of P for ten more steps; then P' for ten more steps; etc. Eventually, during these interlaced simulations, P'' will detect a halt, and when it does so P'' itself halts, returning the appropriate output string.

Here is a program that computes the palindrome problem. First INPUT $C(zz)$ (difficulty $L(S)$, where S is the string entered). Then dump $C(zz)$ into $C(zzz)$, with the order of the characters reversed (difficulty $L(S)$). Then use IF $(C(zz) == C(zzz))$ SKIP ... $\{L(S)\}$ to check for palindrome-ness. This program has difficulty function $L(S)$. So, by the discussion above, this program is at least as efficient as every program computing this problem. Note also that this program has difficulty function $\ll$ the step-difficulty for the Turing computation.

Here is a naive program that computes whether or not a string is prime. To compute whether integer m divides integer n (by the usual long-division method) requires $L(m)(L(n)-L(m)+1)$ steps (for we have to multiply m by a digit ($L(m)$ steps) a total number of times given by $(L(n)-L(m)+1)$. So, we merely check whether the integer $n \geq 2$ entered is divisible, in turn, by each of the integers $2, 3, \cdots, \sqrt{n}$. The difficulty function of this program (at most $\sqrt{n}$ runs, each of difficulty not exceeding $(\log n)^2$) is $\leq \sqrt{n}\,(\log n)^2$ (but is not equivalent to this function, for, e.g., the even integers will be disposed of very quickly by this program). It is easy to write programs that are more efficient than this naive one, e.g., by checking first to see if n is a perfect square, and only if this fails looking for factors of n, as above. In fact, there exist [**?**] [**?**] programs (based on very different methods) that are *much* more

efficient than that above.

> **Exercise.** Find a program that computes whether or not a positive integer is a perfect square; and find its difficulty function.

Conjecture. Given any program (difficulty function f) that computes whether or not an integer is prime, there exists another program that computes that problem, whose difficulty function, f', satisfies $f' \leq f$ and $f' \not\sim f$.

We remark that we could have introduced this language, instead of Turing language, right from the beginning, using it, instead of Turing, as the definition of "computable". Had we done so, then the determination of what can be computed would have been considerably simpler, if perhaps somewhat less illuminating.

12 Are There Better Languages?

Recall that our goal is to obtain the simplest possible language that still captures what we hope is a universal notion of "difficulty". The language constructed in the previous section is intended, as we noted, as merely a suggestion. Here, we comment on a few possible alternatives.

What about dispensing with indexed arrays altogether, i.e., replacing the APPEND, DELETE, and IF commands with subroutines 1-3? This would simplify everything, including the difficulty functions. But, we claim, doing so will likely result in a genuine loss of efficiency. Here is an example. Let the input, S, be a sequence of digits, and set $m = L(S)$. [This will be easier to follow if you think of m as being about 1,000,000, so S is written down, say, in book of some 200 pages.] Now set, for $1 \leq x \leq m$, $f_m(x) = x^{\text{digit}(x)} + 1 \bmod(m)$, where digit$(x)$ means the x^{th} digit of S. Thus, $f_m(x)$ is also an integer between 1 and m. The problem is now the following. Let there be given some input string, S. Start with $x = 7$: Then find $f_m(7)$, then $f_m(f_m(7))$, etc, up to a total of m iterations. Report the result. Let us first compute this problem without benefit of indexed arrays. To determine $f_m(x)$, we must i) find digit(x) (m steps, since we must search through S); and then ii) raise x to a small power ($\leq (\log m)^2$ steps, since x contains at most $(\log m)$ digits). So, the difficulty to compute $f_m(x)$ is $\leq m$, and so the total difficulty to compute the problem (which entails computing $f_m(x)$ m times) is $\leq m^2$. But with indexed arrays, we may first dump the characters

of S into individual locations (via subroutine 13), for a one-time difficulty of $m \log m$. But having done this, computing $f_m(x)$ requires only $(\log m)^2$ steps (one log for locating digit(x), one log for taking the power). This yields a final difficulty function of $m(\log m)^2$. Thus, using indexed arrays is much more efficient than not. The idea of this example is that computing this problem requires that we repeatedly find characters in S, and things are so arranged that which character is to be found is almost random, making it, apparently, impossible to do all the "finding" on a single pass or two through S. It thus becomes more efficient to dump the characters of S into an array, once and for all at the beginning: The resulting easy access to the characters of S ultimately pays off. Of course, we have not *proved* that there exists no way to compute this problem, without indexed arrays, that is much more efficient than the way above, although this looks unlikely. So, the critical issue here is whether our intuitive sense is that the difficulty of this problem should be m^2, or $m(\log m)^2$. If it is the latter, then we must retain indexed arrays.

Even if we begin with commands 1-2 and subroutines 1-3, we could still recover indexed arrays in a simpler way: Introduce two additional basic commands, SET $C(S) = C(C(S'))$ and SET $C(C(S')) = C(S)$. These would allow us to transfer strings currently in indexed arrays to regular locations for further processing, and then to transfer the results back again to the indexed array. What difficulty shall we assign to these commands? We might use $L(C(S')) * L(C(C(S'))$, the difficulty of current subroutine 14. If we do this, then the new language will, apparently, be less efficient than the old. If, for example, we merely want to deal with the last character of a string in an array, $C(C(S'))$, then the original language permits this in just $L(C(S'))$ steps (lookup charge only), while the new language requires that the entire string be copied into a regular location before its last character is accessed. We could avoid this by making the difficulty, for the two new commands above, just $L(C(S'))$. But then the new language would be *more* efficient than the old, for we could copy an entire string from one regular location to another in just 1 step — by copying to an indexed location, and then back. Again, the issue here is what we would like our difficulty function to be.

These complications are caused by lookup charges. Then why not eliminate them entirely, i.e., imagine a world in which looking something up is free, but charges are still made for printing and erasing? This could be achieved, e.g., by retaining the present five classes of commands, but changing the diffi-

culties for each of the last three classes to one. Consider, in this version, subroutine 15. Its difficulty will now be $L(C(C(S'))) * L(C(C(C(S'))))$. Thus, a lookup charge has crept back in: It is reflected in the factor $L(C(C(S')))$, which arises from the necessity to store the string $C(C(S'))$ in order to implement this subroutine. It seems unnatural to have a lookup charge in this case but not in others. We could eliminate that charge here with a new basic command: SET C(S) = C(C(C(S'))). $\{L(C(C(C(S'))))\}$. But then how will we deal with SET $C(S) = C(C(C(C(S'))))$? Again, there will arise a lookup charge if this is made a subroutine, rather than an additional basic command. Are there examples in which such exotic indexed arrays actually impact the final difficulty functions?

Here is a more systematic method by which we might find a natural language with a natural difficulty function. We introduce *machine language(2)*, as follows. Storage locations are labeled by strings of exactly two characters, and each such location always contains exactly one character. Thus, "$C(h8)$" denotes the character in location $h8$; while "$C(C(h8)C(21))$" denotes the character in the location described by the two-character string whose first character is $C(h8)$ and whose second character is $C(21)$. In this machine language(2) there are (in addition to INPUT, OUTPUT, with which we are not concerned right now) four commands:

1. SET $C(xy) = z$.
2. SET $C(C(xy)C(zw)) = C(pq)$
3. SET $C(pq) = C(C(xy)C(zw))$
4. IF $(C(xy) == z)$ SKIP N LINES

where x, y, z, w, p, and q are to be replaced by arbitrary explicit characters, and n by an arbitrary (positive or negative) explicit integer. You can convince yourself that this is enough to carry out simple computations: manipulate strings (whose characters are now stored in individual locations), utilize indexed arrays, branch, count, etc. Indeed, machine language(2) is the *actual* machine language of my (very) old Apple II+. There are 256 characters; and, thus, the total RAM of the computer is just over 65 KB! The good news about machine language(2) is that there is an obvious choice of what difficulty to assign to each command: One step. The bad news is that machine language(2) cannot compute any problem at all (as we have defined

those terms), for it utilizes a finite total memory. You can make available more memory by passing to machine language(3) — the same as that above, except that now *three* characters are needed to describe a location, with the obvious modifications of the basic commands above — or, if still more space is needed, to machine language(4), etc.

The idea, now, is the following. We would introduce a certain basic language, much like that of the previous section; together with a compiler, which would compile programs written in that language into machine language(n) for some n. [Indeed, this is what the Apple II+ does: Here, $n = 2$, and the basic language is Basic.] Given an input string S, the program that is actually run would be the compiled one, written in machine language(n). In this way, we obtain an unambiguous count of steps. If, in the course of that run, it emerged that more memory was needed, then the compiler would kick in again, to recompile the basic-language program in machine language(n') for some $n' > n$. Computation in machine language would then continue. Best if these recompilations could take place seamlessly, e.g., if the machine-language commands could be adjusted so as to be n-universal. Thus, we are free to introduce any sorts of exotic commands we wish in our basic language — the only burden being that these be compiled into machine language. And, we needn't make hard choices as to what the difficulties of these commands are to be: They are whatever follows from their execution in machine language. Thus, since it is the machine language that assigns the difficulties, we might hope that those assignments will be the natural ones. Of course, it would still be required that we decide how to compare number of steps as carried out by machine language(n) with number as carried out by machine language(n'), for $n' \neq n$. It might be interesting to see if this scheme could be implemented.

In any case, let us imagine that there has been introduced a natural language, which gives meaning to "algorithm"; as well as an assignment of difficulty to each command in that language, which gives meaning to "difficulty of that algorithm".

Perhaps the major challenge in this subject is to obtain good lower limits on the difficulty functions for computing various problems. That is, one would like to have theorems of the form: "There is no program that computes this problem π and has difficulty function f satisfying $f \ll$ (something)." Here, "something" is some explicit difficulty function of interest.

Consider, as an example, multiplication of integers. The elementary mul-

tiplication that we all learned in school has difficulty function $\tilde{f}(S) = mn+1$, where m and n are the numbers of digits of the two numbers. [This follows, since, in the course of the multiplication, each digit of the first number must be multiplied by each digit of the second.] Note that there do exist programs that multiply integers, with difficulty function $f \leq mn+1$, and $f \not\sim mn+1$. [Such a program, for example, might first check to see if both integers are integral powers of ten, in which case it writes the product immediately; otherwise, it multiplies the numbers in the usual way.] But does there exist a program that computes this problem, with difficulty function $f \ll mn+1$? It turns out that there does ([**?**]). It is an open question, as far as I am aware, whether there exists a program (with difficulty function f) for multiplication of integers, such that there exists no program with difficulty function $f' \ll f$.

A more famous example is the prime problem, the problem that, given an integer n, returns the prime factors of that integer. The naive program that computes this problem (by trying each integer up to $n^{1/2}$ to see if it divides n) has difficulty just over $n^{1/2}$. It is known that there are much more efficient programs. But what we don't have is a theorem that sets a good lower limit on the difficulty for *any possible* computation of this problem.

This key fact — that we lack good lower limits on the difficulty of computing various problems — has far-reaching implications for the structure of this subject. For instance, as we shall see later, there are examples of problems for which the use of quantum mechanics seems to allow a very efficient computation. In these examples, in particular, it appears that quantum mechanics is more efficient than any known regular computation of that problem. But we cannot *prove* that quantum mechanics is more efficient, for we cannot eliminate the possibility that there exists some — miraculously efficient — regular computation that, for some reason, we have not yet discovered.

13 Probabilistic Computing

As a prerequisite to our study of quantum-assisted computing, we consider here briefly the case of an ordinary computer that has access to a "random number generator". That is, we consider computing in a context in which the actions of a computer, at various stages during its operation, are subject to probabilities. Our purpose here is merely to understand how computing works in this environment. This will allow us, later, to separate effects due

to the full structure of quantum mechanics from those arising solely from its probabilistic character.

Consider the following programming language. The commands are precisely the six introduced in Sect. 11, except for the following change. The third command (APPEND) is replaced by

3. APPEND $x, y, \cdots, z$ TO $C(S)$ $\{1\}$.

Here, $x, y, \cdots z$ stand for any finite list (possibly with repetitions) of explicit single characters from our character set; and, as before, S stands for any explicit string. A program in this new language is defined just as as before (i.e., as a finite list of commands, beginning with an INPUT and ending with an OUTPUT). Whenever, during the running of such a program, one of these new APPEND commands is reached, then whatever string is stored in location $C(S)$ is to be replaced by that same string, but with one of the characters $x, y, \cdots, z$ appended on the right. Which character is appended is to be selected randomly, i.e., with equal probability for each of the characters in the list. Thus, if there are $n \geq 1$ characters in the list, then each of those characters has probability of $1/n$ of being appended. Except for this one change, the programs run just as before. Note that our earlier programming language is a special case of this one, namely, that in which each APPEND command involves a list of exactly $n = 1$ character. We shall call this new language, and its programs, "probabilistic", when we wish to distinguish them from the original language and its programs. This new language is in the spirit of Turing machines and of the old language: We introduce the minimum that is necessary to get the job done.

Clearly, by allowing repetitions of characters in our list, we can achieve any rational-valued probability distribution for which character is appended to $C(S)$. And, combining this new command with the others, we can achieve any rational probabilities for deleting (as opposed to not deleting) a character from the string in any given storage location; as well as rational probabilities for skipping various numbers of lines. Why *rational* probabilities? Why do we not simply allow arbitrary probabilities for appending various characters? This puts us on dangerous ground. Suppose, for example, that we allowed a command that appends "x" to a string, with probability c (the non-computable number of Sect. 7); and appends "y" with probability $1 - c$. Armed with this command, we could write a probabilistic program to com-

pute (in a sense we shall make precise in a moment) the halting problem! In short, we use only rational probabilities in order to prevent sneaking unauthorized information into the program through exotic choices of the probability numbers.

So, let us fix a probabilistic program, and an input string. What can be the result if we run this program on this string? The possibilities in this case are precisely the same as before: The program can halt, with some output string; or it can continue forever without halting. But now, of course, different runs (with the same program and input string) can give different results. Denote by $\tilde{\mathcal{S}}$ the set consisting of all strings over our original character set, $\mathcal{C}$, together with one additional element "$*$", which we designate "not halt". Then we can describe the running of a given program on a given input string by means of a probability distribution on $\tilde{\mathcal{S}}$. That is, for each $\alpha \in \tilde{\mathcal{S}}$, we have a nonnegative number $p(\alpha)$, called the "probability of outcome α", and these satisfy $\sum_{\alpha\in\tilde{\mathcal{S}}}\ p(\alpha) = 1$.

The following example will illustrate these ideas.

> **Example.** Consider the program begins by flipping a coin (i.e., applying an APPEND command with $n = 2$). If the coin comes up "heads", the program reports the total number of coin-flips it has carried out (in this case, "1"), and halts. If the coin is "tails", the program flips the coin again. Again if "heads" comes up, it reports the total number of flips (now "2"); if "tails" it flips again. The program continues in this way. The possible outcomes in this example are the positive integers, together with "$*$". The probability distribution is: For n a positive integer, $p(n) = 2^{-n}$; and $p(*) = 0$.

Note that, in this example, we have $p(*) = 0$ even though it is possible that a given run of this program will never halt. It turns out, however, that this phenomenon can occur only for this special outcome: We claim: For any $\alpha \neq *$, $p(\alpha) > 0$ if and only if α is a possible outcome of running the program. The "only if" is immediate. For "if", let $\alpha \neq *$ be a possible outcome. This means that there exists a sequence of allowed steps in our program that ends with the program halting, with output string "α". There must be only a a finite total number of steps in this sequence (since the sequence ends up with a halt), and so a finite number of APPEND-steps. Let r denote the

(rational) number that results from multiplying the probabilities associated with the given passage through each of these APPEND-steps. Then, clearly, $p(\alpha) \geq r > 0$. **Exercise:** Find an example of a probabilistic program such that the probability of some outcome (say, halting, with output the empty string) is the non-computable number c of Sect. 7.

Now fix a probabilistic program, and also a problem, $\mathcal{S} \xrightarrow{\pi} \mathcal{S}$. We say that this program *(probabilistically) computes* problem π provided that, for every string S, the probability distribution resulting from running this program on initial string S satisfies the following: The probability of failing to halt is zero; and the probability of output string $\pi(S)$ is greater than the probability of every other output string. Think of this definition as requiring that we can extract $\pi(S)$ by "repeated running of the program on S". [We shall make this more precise shortly.] Note that $p(\pi(S))$ can be very small: We only require that no other individual string have probability greater than or equal to that of $\pi(S)$.

Clearly, every (non-probabilistically) computable problem is also probabilistically computable, since every non-probabilistic program is already a probabilistic program (namely, one in which each APPEND command happens to have but a single choice). It turns out that the converse is also true: Every probabilistically computable problem is also (non-probabilistically) computable. In other words, the introduction of probability adds nothing to what can be computed.

To prove this, fix a problem π, and a probabilistic program, $\mathcal{P}$, that computes it, in the sense above. We now construct non-probabilistic program, $\tilde{\mathcal{P}}$, as follows. Given any input string S_o, this $\tilde{\mathcal{P}}$ simulates the action of $\mathcal{P}$ on S_o. That is, $\tilde{\mathcal{P}}$ keeps track, at each step, of which program line $\mathcal{P}$ is currently executing; and what string resides in each of $\mathcal{P}$'s storage locations. Then $\tilde{\mathcal{P}}$ simply follows the action of $\mathcal{P}$, step by step. When $\mathcal{P}$, so simulated, reaches an APPEND command, there will in general be several options for the next state (corresponding to the several possible characters that could be APPENDed in response to this command). When this happens, $\tilde{\mathcal{P}}$ simply keeps track of each of these options separately, and also keeps a record of the probability for each. Thus, for example, if the simulation by $\tilde{\mathcal{P}}$ reaches the command APPEND x, y, z TO $C(k8)$, then $\tilde{\mathcal{P}}$ will consider separately the cases in which x, or y, or z is appended to the string $C(k8)$, assigning probability 1/3 to each. Then $\tilde{\mathcal{P}}$ will simply simulate the action of $\mathcal{P}$ separately for each of the three cases. This branching continues for subsequent APPEND

commands: If, say, one of these branches reaches another APPEND command, then there will result further branches (with new probability assignments) for $\tilde{\mathcal{P}}$ to follow.

Now, as $\tilde{\mathcal{P}}$ continues to follow all these branches, there will occur, every so often, a branch on which $\mathcal{P}$ would have encountered an OUTPUT command, and thus would have halted. When $\tilde{\mathcal{P}}$ reaches this point of a branch, then, of course, it can no longer follow that branch, since $\mathcal{P}$ itself would be unable to continue to operate along that branch. The program $\tilde{\mathcal{P}}$ maintains a table in which there is recorded, for each such terminated branch, two pieces of information: The $\mathcal{P}$-output string at that $\mathcal{P}$-halt, and the probability (a rational number) of reaching that particular termination. As $\tilde{\mathcal{P}}$ continues its simulation new terminating branches will be found, and this table will continue to grow. Note that the sum of the probabilities in this table will always be less than or equal to one. It further follows, from our assumption that $p(*) = 0$, that this sum will approach one, as $\tilde{\mathcal{P}}$ continues to run in this way.

Each time the program $\tilde{\mathcal{P}}$ adds a new line to this table, it will also perform the following calculation. First, it finds that string, S, having the largest total probability (i.e., that string such that the sum of the probabilities already assigned to that string in the table is greater than the sum of the probabilities already assigned to any other string). Then, $\tilde{\mathcal{P}}$ computes how much probability remains, i.e., it computes the number given by subtracting, from one, the sum of all the probabilities listed in the table. Next, $\tilde{\mathcal{P}}$ asks: Is there any other string, S', such that, if all the remaining probability were allocated to S' (in addition to the probability already assigned to S'), then this total would exceed the probability for S? If the answer to this question is yes, i.e., if there does exist a string S' having the potential of ultimately accumulating more probability than has already been assigned to S, then $\tilde{\mathcal{P}}$ continues to run. But eventually $\tilde{\mathcal{P}}$ must reach a point at which the answer to this question is no. That is, it will reach a point at which some string S has already accumulated enough probability that no other string is even a candidate ever to accumulate more. [This follows from the fact that the program $\mathcal{P}$ computes the problem π, i.e., that $p(*) = 0$ and $p(\pi(S_o))$ exceeds the probability of every other string.] When this happens, $\tilde{\mathcal{P}}$ itself halts, and announces the winning string, S.

Clearly, this (non-probabilistic) program $\tilde{\mathcal{P}}$ also computes the problem π. What we have shown, then, is that, given a probabilistic program that

probabilistically computes a problem, we can, using simulation, build a non-probabilistic program that computes the same problem[5]. In short, the use of probability adds nothing to the concept of computability.

We now turn to the issue of assigning a difficulty function to a probabilistic computation.

Fix a probabilistic program, $\mathcal{P}$, that computes a problem π (so, for any input string S, the probability that $\mathcal{P}$ fails to halt for this string is zero; and the most likely output string is $\pi(S)$). Fix any input string S, and let us run the program $\mathcal{P}$ on that string. Then during this run, various commands will be executed, and to each of these we have assigned a difficulty. Let us keep track of the cumulative total difficulty during the running of the program. Now should it happen, on this particular run of $\mathcal{P}$, that the program fails to halt, then the cumulative difficulty will, of course, grow without bound. But if $\mathcal{P}$ does halt, then there will be some total cumulated difficulty, ν, as of that halt. On different runs, there will be different cumulated difficulties. Thus we shall have some probability distribution on the possible cumulated difficulties, i.e., for each ν, we have a number $p(\nu) \geq 0$, such that $\sum_\nu p(\nu) = 1$. [That this sum must actually be one follows from $p(*) = 0$.] Denote by $D(S)$ the mean total difficulty: $D(S) = \sum_\nu \nu\, p(\nu)$. This $D(S)$ is the difficulty that would be experienced "on the average" in one run of $\mathcal{P}$ with the given input string S. Of course, it is only an average: On any given run, it is entirely possible that the actual cumulated difficulty turn out to be much greater than $D(S)$ — or much less. Note that the sum defining $D(S)$ need not converge: The difficulty ν could grow very quickly even as $p(\nu)$ approaches zero. [**Exercise:** Find an example.] If this should occur, then we assign $\mathcal{P}$ infinite difficulty for the input string S, and abandon further efforts to assign a difficulty function for this program. Note that, by simulating the running of $\mathcal{P}$ on input string S, as described above, we could compute an increasing sequence of rational numbers that converges to $D(S)$ (or, in the case in which $D(S) = \infty$, that grows without bound). It seems unlikely, nevertheless, that D is always computable, in the sense that there always

[5]In fact, this remains true even with certain, even weaker, notions of "probabilistically computable". For example, it suffices to require, instead of $p(*) = 0$, merely that we are given a program that, for each input string, computes the number $p(*)$. We also remark that one can modify this simulation program $\tilde{\mathcal{P}}$ to prove the following: If, for any probabilistic program acting on any string, $p(*)$ is computable, then each of the probabilities for each of the possible output strings is also a computable number.

exists a (regular) program that, given probabilistic program $\mathcal{P}$, string S, and a positive integer n, returns a rational within $1/n$ of $D(S)$. Indeed, even the problem of whether or not $D(S)$ is finite is probably not computable.

In any case, we have the notion of the mean difficulty, $D(S)$, for one run of $\mathcal{P}$ with input string S. But, unfortunately, a single run of this program, with input string S, does not tell us what the answer to our problem π is for that string, i.e., does not tell us what $\pi(S)$ is. Rather, we must run the program a number of times, on the same given input string, and keep a record of the various outputs. The "real" answer will be buried in the statistics of these records (in the form of the "most likely" output). What me must determine, then, is by what factor to multiply the mean difficulty, $D(S)$, to correct for this probabilistic character. To this end, let us run this program a total of r times, keeping a record of the various outputs that result. At the end of all these runs, we announce as the answer that output that occurred most frequently. Sometimes we will announce the correct answer, $\pi(S)$, and sometimes the wrong answer. Denote by $\kappa(r)$ the probability that our announcement is wrong. The following lemma states, roughly speaking, that, as the number r of runs increases, this probability $\kappa(r)$ goes to zero as e^{-Kr}, for a certain number K:

Lemma. Consider a collection of positive numbers, with sum one. Denote the largest by p and the next largest by p', and assume $p > p'$. Carry out r runs in the corresponding probability distribution, and denote by $\kappa(r)$ the probability that the most frequent single outcome is not the most probable outcome (i.e., not the p-outcome). Then the limit of $[-\log \kappa(r)/r]$, as $r \to \infty$, exists, and has value $K = (p-p')^2/2[p(1-(p-p'))^2+p'(1+(p-p'))^2]$.

The proof uses three facts: i) For large r, any other outcome, say with probability $p'' < p'$, has negligible probability (compared with that of p') of being the most frequent outcome; ii) the difference between the numbers of p-outcomes and p'-outcomes is, for large r, normally distributed, with mean $r(p-p')$ and squared-variance $r[p(1-(p-p'))^2 + p'(1+(p-p'))^2]$; and iii) the error function, $\mathrm{erf}(x)$, satisfies $\lim_{x\to\infty}\mathrm{erf}(x)/x^2 = -1/2$. For the present application, the p of the lemma is $p(\pi(S))$, and the p' is the probability of the next-most-likely outcome. Note that the number K of the lemma here depends on the input string S (through the dependence of the probabilities p, p' on S).

Now fix a small number $p_o > 0$, which we shall interpret shortly as a confidence limit, i.e., as the "largest probability of error that we are willing to tolerate in our determination $\pi(S)$". Fix the input string S. Let us now run our program r_o times, keeping track of the outputs for each run, and report as the answer that outcome that occurs most frequently in these runs. We wish to choose r_o sufficiently large that the probability that this procedure results in the wrong answer does not exceed our confidence limit p_o. It follows from the lemma that, at least for sufficiently small p_o, the choice $r_o \geq -(\log p_o)/K$ suffices, where K is the expression given in the lemma. That is, carrying out $r_o \geq -(\log p_o)/K$ runs and reporting the most frequent outcome will, with probability at least $1 - p_o$, result in reporting $\pi(S)$. Note that, as we expect, the number of runs required grows without bound as $p_o \to 0$.

We now take, as the difficulty of computing $\pi(S)$ using the probabilistic program $\mathcal{P}$ on input string S, the number $[-(\log p_o)/K]D(S)$, i.e., the product of the minimum number (r_o) of runs required and the mean difficulty ($D(S)$) per run. Repeating this procedure for all possible input strings S, we obtain the difficulty function $f(S) = -(\log p_o)D(S)/K$ for this probabilistic computation of the problem π. But note that the confidence limit p_o appears only in an overall factor. Thus, up to equivalence, it may be omitted. That is, it makes no difference how small is the confidence limit p_o we choose: The resulting difficulty function, up to equivalence, is independent of p_o. There results $f(S) = D(S)/K$. But this expression is not as it stands suitable for a difficulty function, because it is not in general bounded away from zero. [Indeed, as $p \to 1$ (whence $p' \to 0$), $K \to \infty$.] The reason for this phenomenon is quite simple: The argument above requires, in the limit $p \to 1$, that the program $\mathcal{P}$ be run only a small fraction of one time! But at least one full run of $\mathcal{P}$ is necessary in any case, and we can take this fact into account by adding $D(S)$ to the $f(S)$ above. Doing this, and passing to an equivalent difficult function, we obtain

> Let $\mathcal{P}$ be a probabilistic program that computes some problem, π. Then we shall assign to this program the difficulty function given by $f(S) = D(S)(p + p')/(p - p')^2$, where $D(S)$ is the mean difficulty for running $\mathcal{P}$ on input string S, p is the probability that that run results in output $\pi(S)$, and $p' < p$ is the probability of the next most probable output.

The factor, $(p+p')/(p-p')^2$, by which $D(S)$ is multiplied reflects the increase in difficulty due to the fact that $\mathcal{P}$ computes our problem only probabilistically. This factor is always at least one, and for $p = 1$ (and so $p' = 0$) this factor is exactly one, i.e., the difficulty function for a probabilistic program reduces, in the special case of a non-probabilistic program, to our original difficulty function. When p and p' are very close, the factor is large, reflecting the fact that there must be carried out many runs of $\mathcal{P}$, on the given input string S, in order to have a reasonable chance of announcing the correct value of $\pi(S)$.

We have already seen that the introduction of probability adds nothing to what is computable. But can probability add to efficiency? Consider

> **Assertion:** Let $\mathcal{P}$ be a probabilistic program that probabilistically computes problem π, with difficulty function f. Then there exists a non-probabilistic program $\mathcal{P}'$ that also computes π, and whose difficulty function f' satisfies $f' \leq f$.

This assertion states, in other words, that any probabilistic program can always be at least matched, in terms of efficiency, by a corresponding non-probabilistic program. It seems likely, intuitively, that this assertion is true. Given string S, there is some definite string, $\pi(S)$, that you wish to compute. Why would it ever be more efficient to use a random-number generator to find this $\pi(S)$?

One might imagine that one could prove this assertion by using a simulation, as described earlier. Given probabilistic program, $\mathcal{P}$, that computes π, then we can, by simulating $\mathcal{P}$, construct a non-probabilistic program, $\tilde{\mathcal{P}}$, that also computes π. If we could show that the difficulty of $\tilde{\mathcal{P}}$ is always less than or equal to that of $\mathcal{P}$, then we would be done. Unfortunately, this is false in general.

> Let π be the problem that assigns to each string (represented as a positive integer) the string "a". Consider the probabilistic program $\mathcal{P}$ that operates as follows. Given input n, $\mathcal{P}$ flips a coin n times, and then simply rolls a die. If the die comes up "1", $\mathcal{P}$ reports "b"; otherwise, $\mathcal{P}$ reports "a". This $\mathcal{P}$ computes the problem π, and has difficulty function f given by $f(n) = n$. Denote by $\tilde{\mathcal{P}}$ the non-probabilistic program that simulates $\mathcal{P}$, as

described earlier. Thus, $\tilde{\mathcal{P}}$ also computes π. But, because of the 2^n branches created by $\mathcal{P}$ (via its coin-flips), this $\tilde{\mathcal{P}}$ has difficulty function $\tilde{f}(n) = 2^n$. Thus, $f \ll \tilde{f}$.

Thus, the program $\mathcal{P}$ merely creates 2^n branches, and then proceeds to ignore them! Clearly, there is no need, in this example, for $\tilde{\mathcal{P}}$ to follow all 2^n branches: They are all the same, so it suffices for $\tilde{\mathcal{P}}$ to follow a single branch. It is obvious, in this example, how $\tilde{\mathcal{P}}$ can avoid unnecessary computations. But in order to prove the assertion we need to find a general way for $\tilde{\mathcal{P}}$ to do this. For example, the probabilistic program $\mathcal{P}$ could actually ignore many branches, but could be so written to disguise this fact from $\tilde{\mathcal{P}}$. Thus, if such a simulation is ever going to work to produce a proof of the assertion, then it will be necessary to design a "smart" simulation program $\tilde{\mathcal{P}}$ — one that looks ahead to find what are the more promising branches. That this idea can be implemented is by no means obvious. Thus, it is not at all clear whether or not there exists a problem, and probabilistic program that computes that problem, which stands as a counterexample to the assertion above.

Amazingly enough, this assertion remains open! Even an example of a problem, together with a probabilistic computation of that problem, such that it appears plausible that there is no non-probabilistic computation that is at least as efficient, would be most interesting. Here is an possible strategy to obtain such an example.

Fix a problem π_o that accepts as input any pair of positive integers, (N, k), with $k \leq N$, and produces as output either "yes" or "no". Thus, for each value of N a total of N values of k are allowed, $1, \cdots, N$. Further, let this π_o have the following property[6]: For each N, either i) $\pi_o(N, k) =$ "yes" for at least two-thirds of the allowed k-values, or ii) $\pi_o(N, k) =$ "no" for at least two-thirds of the allowed k-values. From this π_o, we construct a new problem, π, acting on positive integers N, as follows: $\pi(N)$ is "yes" or "no", according as whether case i) or ii) above applies for that N. This π is the problem of interest.

Now let there be given a program, $\mathcal{P}_o$, that computes π_o. Then we can easily write a program, $\mathcal{P}$, that computes π: Given any positive integer N, let the program $\mathcal{P}$ simply run $\mathcal{P}_o$ on (N, k) for each of the allowed k-values, $k = 1, \cdots, N$, and report "yes" or "no" according to which answer resulted

[6]This property is imposed in a very strong form, for ease of exposition. It can be weakened considerably.

more frequently in those N runs. [Actually, it is enough to run $\mathcal{P}_o$ for just over two-thirds of these k-values.]

Denote by f_o the difficulty function of $\mathcal{P}_o$. For each positive integer N, denote by $g(N)$ the maximum value achieved by $f(N,k)$, as k runs through $1, \cdots, N$. Let us also suppose that, for each fixed N, all of the $f_o(N,k)$, for $k = 1, \cdots, N$, are between, say, $g(N)/2$ and $g(N)$. [This supposition is made merely to simplify the discussion: It could be weakened considerably.] Then, up to equivalence, the difficulty function of the program $\mathcal{P}$ (since it merely runs $\mathcal{P}_o$ a total of N times) is $Ng(N)$.

Here is a probabilistic program, $\mathcal{P}_{\text{prob}}$ that also computes the problem π. For each positive integer N, let $\mathcal{P}_{\text{prob}}$ select, randomly, a positive integer $k \leq N$, run the program $\mathcal{P}_o$ on the pair (N,k), and report whatever $\mathcal{P}_o$ reports on this single run. Note that this probabilistic program $\mathcal{P}_{\text{prob}}$ always halts; and that, for each N, the probability of its giving the correct answer is at least two-thirds. Thus, this $\mathcal{P}_{\text{prob}}$ does indeed probabilistically compute π. The probabilistic program $\mathcal{P}_{\text{prob}}$ has difficulty function given simply by $g(N)$. That is, the probabilistic program $\mathcal{P}_{\text{prob}}$ is *far* more efficient than the non-probabilistic program $\mathcal{P}$.

So, it would appear to be relatively easy to find an example of a problem for which the probabilistic program is more efficient than the non-probabilistic. All we must do is find a problem π_o of the type described above. Here is a simple example: Let $\pi_o(N,k)$ be "yes" if k is a prime; and "no" if k is composite. Then this π_o (with suitable adjustments for the first few N-values) satisfies the "two-thirds"-condition above. In this example, then, for given N, the non-probabilistic program $\mathcal{P}$ must check for primeness of each integer $k = 1, \cdots, N$; while probabilistic program $\mathcal{P}_{\text{prob}}$ checks for primeness of single (randomly selected) integer in this range. Clearly, $\mathcal{P}_{\text{prob}}$ is far more efficient than $\mathcal{P}$. But in this case there exists a shortcut for computing the problem π — there exists a non-probabilistic program that is much more efficient than either of these two. This is the program that (except possibly for the first few N-values) always reports "no". Thus, the probabilistic program $\mathcal{P}_{\text{prob}}$ in this example, while more efficient than $\mathcal{P}$, is not more efficient than *every* non-probabilistic program that computes π. This example, in other words, fails.

What is needed, then, is a computable problem π_o that reports, for each N, either "yes" for at least two-thirds of the allowed k-values $(1, \cdots, N)$ or "no" for at least two-thirds of the allowed k-values — but which is such

that there is no shortcut for determining which of these answers is the more frequent. It must be the case that the only way to determine whether two-thirds of the answers are "yes" or two-thirds "no" is actually to compute $\pi_o(N, k)$ for the requisite number of k-values (or, at least, to do some equally difficult computation). Remarkably enough, no such example of a π_o seems to be known! A theorem to the effect that there exists no such π_o — i.e., a theorem to the effect that this type of example will always fail — would be extremely interesting.

For most probabilistic programs of interest, the probability of the correct outcome dominates the other probabilities, and when this is the case the formula above can be simplified. Suppose that there exists a number $a > 0$, independent of S, such that $p(\pi(S))$ exceeds the next-highest probability by at least this a. This condition is always satisfied, e.g., if all the $p(\pi(S))$ are greater than 0.501. Indeed, one could take the position that $\mathcal{P}$ is not "really computing" the problem unless this condition is satisfied. In any case, whenever this condition is satisfied, then the factor $(p + p')/(p - p')^2$ in the formula above is bounded above, and so in this case the difficulty function, up to equivalence, is given simply by $D(S)$. That is: Under this rather weak condition, we may assign to a program $\mathcal{P}$ that solves the problem π the difficulty function whose value on any input string is the mean difficulty of running that program on that input string.

> **Exercise.** Let $\mathcal{P}$ and $\mathcal{P}'$, with respective difficulty functions f and f', both compute the same problem. Is there a way to alternate between these two programs, constructing a program $\mathcal{P}''$ that also computes this problem, with difficulty function given by $f''(S) = \min(f(S), f'(S))$?

The probabilistic programs that one would typically write will have the property that, for every run of the program on every input string, the program will always halt. [As we have seen from an earlier example, this property is stronger than merely requiring that $p(*) = 0$.] An example would be a Monte-Carlo program: It visits an APPEND command with several outcomes a certain number of times; keeps track of the results, thus generating a distribution of outcomes; and then simply halts, reporting, say, some property of this distribution. For such a program — one that always halts, on every run with every input string — the discussion above can be simplified considerably. Fix such a program, $\mathcal{P}$, and an input string S. We claim, first, that,

in this special case, there is but a finite number of possible outcomes. To see this, call a state, during the running of this program, *rich* if there is an infinite number of possible outcomes starting from that state. Suppose, for contradiction, that the initial state — the original INPUT of the initial string — were rich. Now follow the steps of this program as it runs. As long as we encounter only non-APPEND commands, we must always remain in a rich configuration. Now consider the first APPEND-command we encounter. We are already in a rich state at this point, and so at least one of the possible branches from this point must itself be rich. Choose any such rich branch. Continuing in this way (taking, at each probability APPEND-command, some choice that again results in a rich configuration) we will always remain in a rich configuration; and therefore we can never halt (since the halt-state is hardly rich). But this contradicts our assumption that failure to halt, for any run on any string, is not possible. This shows that our initial supposition — that there was an infinite number of possible outcomes — must be false. A similar argument shows that, in this situation (a program, running on an input string, such that the failing to halt is not a possible outcome), there must be an upper bound to the number of steps that will ever be required to achieve the halt. The proof is the identical to that above, merely redefining "rich" as "having no upper bound for the maximum number of steps that could be required, from that point, to achieve the halt". These two proofs are essentially the same as that of the "tree theorem" in mathematics. It also follows, in this case, that the probabilities for the different output strings are all rational (since there is a finite number of possible routes to a given output string, and each of these, since it encounters a finite number of APPEND commands, has a rational probability of being the actual route). And finally, the simulation program $\tilde{\mathcal{P}}$ that we introduced earlier will, in this case, halt all by itself (without invoking the special rule involving a probability calculation). This follows, since each of the branches in the simulation must, eventually, be terminated.

Thus, many of the complications of simulating the running of a probabilistic program, and of computing its difficulty function, disappear in the special case that the program, for every run with every input string, always halts.

14 Quantum Mechanics

This section is a very short course in quantum mechanics — for people who already know quantum mechanics.

A *Hilbert space* is a complex vector space, equipped with an inner product that is antilinear in the first factor and linear in the second, such that the associated norm is positive-definite. All our Hilbert spaces will be finite-dimensional/footnote The full definition of a Hilbert space includes an additional condition of completeness, but in the finite-dimensional case completeness follows automatically. Vectors in Hilbert spaces are usually written, e.g., as $|\alpha\rangle$, where α is some symbol or word that describes the vector; and the inner product of vectors $|\alpha\rangle$ and $|\beta\rangle$ is usually written $\langle\alpha|\beta\rangle$. *The states of a quantum system are described by nonzero vectors (up to an overall complex factor) in a suitable Hilbert space.*

Let H and H' be Hilbert spaces. The *tensor product* of H and H' is a certain Hilbert space obtained by taking linear combinations of formal products, where each product is of one vector in H with one vector in H'. The tensor product is written $H \otimes H'$, and has dimension given by the product of the dimensions of H and H'. For $|\alpha\rangle \in H$ and $|\alpha'\rangle \in H'$, the corresponding formal product, in $H \otimes H'$, is written $|\alpha\rangle|\alpha'\rangle$. *Now consider two quantum systems, whose states are described by respective Hilbert spaces H and H'. Regard these two separate systems as one. Then the Hilbert space of states of the combined system is $H \otimes H'$. Indeed, $|\alpha\rangle|\alpha'\rangle$ represents that state of the combined system with the H-system in state $|\alpha\rangle$ and the H'-system in state $|\alpha'\rangle$. Since the Hilbert space $H \otimes H'$ allows linear combinations of these simple products, not every state of the combined system is one in which each of the original systems is in a particular state.*

An *operator* on a (finite-dimensional) Hilbert space H is a linear mapping from H to itself. For example, the identity, I, is an operator, as is, for any $|\alpha\rangle \in H$, the map, written $|\alpha\rangle\langle\alpha|$, with action $|\alpha\rangle\langle\alpha|\ (|\beta\rangle) = |\alpha\rangle\ (\langle\alpha|\beta\rangle)$. For A an operator and $|\alpha\rangle$ a vector in the Hilbert space, we sometimes write $|A\alpha\rangle$ for $A(|\alpha\rangle)$. For A and A' operators on Hilbert spaces H and H', respectively, we write $A \otimes A'$ for the operator on $H \otimes H'$ with action $(A \otimes A')(|\alpha\rangle|\alpha'\rangle) = |A\alpha\rangle|A'\alpha'\rangle$ (extended to all of $H \otimes H'$ by linearity). We shall sometimes not distinguish between an operator A' acting on H' and the operator $I \otimes A'$ acting on $H \otimes H'$.

An operator U on a Hilbert space is called *unitary* if it is inner-product

preserving, i.e., if $\langle U\alpha|U\beta\rangle = \langle\alpha|\beta\rangle$ for every α, β. For example, if $|\alpha\rangle$ is unit, then $I - 2|\alpha\rangle\langle\alpha|$ is unitary. *The evolution of a quantum system through time is described by a unitary operator U: Initial state $|\psi\rangle$ evolves to $|U\psi\rangle$.*

An operator A on a Hilbert space is called *Hermitian* if it satisfies $\langle A\alpha|\beta\rangle = \langle\alpha|A\beta\rangle$ for every α, β. For example, I and $|\gamma\rangle\langle\gamma|$ are Hermitian. In the finite-dimensional case, every Hermitian operator has a finite number of eigenvalues, all real, and the corresponding eigenspaces span the entire Hilbert space. *Observations on quantum systems are described by Hermitian operators. Let a system, initially in state given by unit $|\psi\rangle$, be observed via Hermitian A. Then the "result" of the observation is one of the eigenvalues of A; the state of the system after the observation is the projection of $|\psi\rangle$ into the corresponding eigenspace; and the probability of that result is the squared-norm of that projection. Given a basis for H, by an observation* via that basis *we mean an observation via a Hermitian operator whose eigenspaces are those generated by the individual basis vectors.*

15 Grover Construction

We now begin a new subject: quantum-assisted computing. Our strategy will be first to consider, in some detail, one particular example. We shall then generalize. We choose for our example what is called the Grover construction [**?**][**?**][**?**], for it has a number of attractive features: It is very simple; it illustrates most of the constructs and ideas of quantum-assisted computing; and it holds out realistic hope of generating an example in which the quantum-assist provides a genuine reduction in difficulty.

Consider the challenge of finding a needle in a haystack. Fix an integer N (which you should think of as containing, say, 100 digits). The haystack is the N integers $0, 1, \cdots, (N-1)$; and the needle is a specific one of those integers, say k_o. We suppose that we have a computer that allows us to search for the needle in the following manner. The computer accepts as input any integer k with $0 \le k \le (N-1)$, and returns either "no" (if $k \ne k_o$) or "yes" (if $k = k_o$). We wish to find the needle. The obvious way to do this is to run the computer for various k-values as input. Thus, to be certain of finding k_o we would have to run the computer a total of N times; while a mere 50% chance would require only $N/2$ runs. The issue is whether we can discover a way to find the needle in substantially fewer runs.

Here is a corresponding quantum system. Let there be given an N-dimensional Hilbert space, H_{in}, with orthonormal basis $|0\rangle, |1\rangle, \cdots |N-1\rangle$: This is the quantum system in which the input will be registered. And, similarly, let there be given 2-dimensional Hilbert space, H_{out}, with orthonormal basis $|\text{no}\rangle, |\text{yes}\rangle$, to register the output. Then the Hilbert space with which the computer (and we) interact is $H_{\text{in}} \otimes H_{\text{out}}$. We represent the action of the computer by the following unitary operator[7] on this Hilbert space:

$$V(|k\rangle|\text{no}\rangle) = |k\rangle|\text{no}\rangle \ (k \neq k_o) \qquad V(|k_o\rangle|\text{no}\rangle) = |k_o\rangle|\text{yes}\rangle, \tag{4}$$

$$V(|k\rangle|\text{yes}\rangle) = |k\rangle|\text{yes}\rangle \ (k \neq k_o) \qquad V(|k_o\rangle|\text{yes}\rangle) = |k_o\rangle|\text{no}\rangle. \tag{5}$$

That is, if the input register is in any state other than $|k_o\rangle$, then V does nothing; while if it is in state $|k_o\rangle$, then V flips the output state. This unitary operator V is a reasonable rendition of what a computer might do. Indeed, suppose we have agreed to start the system with the output register in state $|\text{no}\rangle$. Then Eqn. (**??**) above specifies that V records the correct answer (for the given $|k\rangle$) in H_{out}. And (**??**) is the simplest way to extend this V, as a unitary operator, to all of $H_{\text{in}} \otimes H_{\text{out}}$.

Let us pause at this point to see how we might search for the needle under this setup. First select any candidate k, then begin with the registers in the corresponding initial state, $|k\rangle|\text{no}\rangle$, and then run the computer (i.e., apply V). When the computer is finished (with final register-state that given in (**??**)-(**??**)), make an observation, on H_{out}, via the basis $|\text{no}\rangle, |\text{yes}\rangle$. If the result is "no" (which it will be, with probability $(N-1)/N$), then we know that our trial k was not the needle; while if it is "yes" (probability $1/N$) then we have found our k_o. This will be recognized as merely the original search, cloaked in a thin veneer of quantum mechanics.

Let us now change things slightly. Set $|\phi\rangle = \frac{1}{\sqrt{N}}(|0\rangle + |1\rangle + \cdots + |N-1\rangle)$, a unit vector in H_{in}. This is a state that combines all possible inputs, equally weighted. Let us now begin with state $|\phi\rangle|\text{no}\rangle$. Then the running of the computer produces

$$V(|\phi\rangle|\text{no}\rangle) = \frac{1}{\sqrt{N}} \ \{|0\rangle + \cdots + |k_o - 1\rangle + |k_o + 1\rangle + \cdots + |N-1\rangle\} \ |\text{no}\rangle$$

[7]The action of V on the linear combinations of these simple product states is, of course, fixed by linearity.

$$+\frac{1}{\sqrt{N}}|k_o\rangle|\text{yes}\rangle.$$

Again, let us see what we can learn from this final state. We first make an observation on H_{out} via its basis. With probability $(N-1)/N$ we will obtain "no", in which case we have learned nothing whatever (not even, as in the previous paragraph, a k known not to be the needle). But, one time out of N, we will get lucky and obtain "yes". In this case, we proceed to make an observation on H_{in} via its basis $|0\rangle, |1\rangle, \cdots, |N-1\rangle$. The result (since now the H_{in}-state is simply $|k_o\rangle$) will tell us what k_o is. But note that even this procedure, using the state $|\phi\rangle \in H_{\text{in}}$, hasn't gained us anything: This is still basically the original search, the only essential difference being that now quantum mechanics is "choosing" our trial k's for us.

Let us now make still another change, this time to the output register. Let us now choose as our initial state $|\phi\rangle\ \frac{1}{\sqrt{2}}\ \{|\text{no}\rangle - |\text{yes}\rangle\}$. In this case, the running of the computer produces

$$V(|\phi\rangle\ \frac{1}{\sqrt{2}}\{|\text{no}\rangle - |\text{yes}\rangle\})$$

$$= \frac{1}{\sqrt{N}}\ \{|0\rangle + \cdots + |k_o - 1\rangle - |k_o\rangle + |k_o + 1\rangle + \cdots + |N-1\rangle\}\frac{1}{\sqrt{2}}\ \{|\text{no}\rangle - |\text{yes}\rangle\}.$$

That is, the output register is now always in the state $\frac{1}{\sqrt{2}}\{|\text{no}\rangle - |\text{yes}\rangle\}$ — both before and after the running of the computer. All the computer does, now, is reverse of the sign of the $|k_o\rangle$-term in the input register. What can we learn by making our observations on *this* final state? Absolutely nothing. An observation on H_{out}, via its basis, will give equal probability for "no" and "yes"; and an observation on H_{in}, via its basis, will return each $k = 0, 1, \cdots, (N-1)$ with equal probability. It looks as though we have gone backward.

Undaunted, we set $V_{\text{in}} = I - 2|k_o\rangle\langle k_o|$, a unitary operator (reflection across the plane orthogonal to $|k_o\rangle$) on H_{in}. Then the result of the previous paragraph can be summarized as follows: For any $|\psi\rangle \in H_{\text{in}}$,

$$V(|\psi\rangle\ \frac{1}{\sqrt{2}}\{|\text{no}\rangle - |\text{yes}\rangle\}) = |V_{\text{in}}\psi\rangle\ \frac{1}{\sqrt{2}}\{|\text{no}\rangle - |\text{yes}\rangle\}.$$

That is, provided the H_{out}-state is set to $\frac{1}{\sqrt{2}}\{|\text{no}\rangle - |\text{yes}\rangle\}$, the action of V (the run-the-computer operator) on $H_{\text{in}} \otimes H_{\text{out}}$ is represented by the action

of this $V_{\rm in}$ on $H_{\rm in}$, the $H_{\rm out}$-state never changing. Next, set $W = I - 2|\phi\rangle\langle\phi|$, another unitary operator (reflection across the plane orthogonal to $|\phi\rangle$) on $H_{\rm in}$. Note that W does *not* involve knowing which $|k\rangle$ is the needle in the haystack. We now have, by an easy calculation,

$$-WV_{\rm in}|\phi\rangle = \frac{N-4}{N}\ |\phi\rangle + \frac{2}{\sqrt{N}}\ |k_o\rangle. \tag{6}$$

Thus, we are now working solely in $H_{\rm in}$, for we begin with $H_{\rm out}$-state $\frac{1}{\sqrt{2}}\{|{\rm no}\rangle - |{\rm yes}\rangle\}$, and this state never changes. Eqn. (**??**) gives the result of starting with state $|\phi\rangle \in H_{\rm in}$, then running the computer (i.e., applying unitary $V_{\rm in}$), and then applying unitary W.

Again, let us pause to interpret this equation. Let us make an observation, on the state given by the right side of (**??**), via our basis, $|0\rangle, |1\rangle, \cdots, |N-1\rangle$, for $H_{\rm in}$. We find (taking the inner product of that right side with $|k_o\rangle$ and squaring the result) that the probability of obtaining k_o is $(3N-4)^2/N^3$ (the rest of the probability being distributed equally over the other k's). For large N, this probability is about $9/N$. After observing via this basis (obtaining a k-value), we may of course check directly, by running our classical computer, whether that k is actually the needle. Nine times out of N, we will in this way find the needle. Note that this is nine times the a priori probability of finding k_o by merely guessing a k-value. It may look as though we are making some real progress here, but this appearance is misleading. Even a factor of nine in the probability for success still means that, in order to find the needle, we must carry out a number of runs proportional to N. But suppose that, instead of observing the state (**??**) immediately, we repeat the operation: Apply $-WV_{\rm in}$ again, and only then observe via the $|k\rangle$-basis and check the k-value that results? Our probability of success will then turn out to be twenty-five times the a priori probability. These remarks motivate what follows.

Now comes the key step: To look, from a geometrical viewpoint, at what we have just done. Consider the 2-plane in $H_{\rm in}$ spanned by $|k_o\rangle$ and $|\phi\rangle$. Each of the operators of interest, $V_{\rm in}$ and W, when acting on any vector orthogonal to this 2-plane, is the identity. Thus, all the action is taking place within this 2-plane. Let us choose an orthonormal basis for this 2-plane consisting of $|k_o\rangle$ and $|k_o\rangle^\perp$, where the latter is that linear combination of $|k_o\rangle$ and $|\phi\rangle$ that is unit and orthogonal to $|k_o\rangle$. Denote by θ the angle that $|\phi\rangle$ makes with $|k_o\rangle^\perp$. Then $\sin\theta = \langle k_o|\phi\rangle = \frac{1}{\sqrt{N}}$.

Now, each of $V_{\rm in}$ and $-W$ is a certain reflection within this plane (about vectors $|k_o\rangle^\perp$ and $|\phi\rangle$, respectively). But the composition of two reflections in a plane is a rotation. The angle of rotation is given by cos(angle) = $\langle\psi|(-WV_{\rm in})\psi\rangle$, where $|\psi\rangle$ is any unit vector in our 2-plane. Choosing $|\psi\rangle = |\phi\rangle$ (or $|k_o\rangle$, if you prefer), we find that this angle is precisely 2θ.

So, vector $|\phi\rangle$ starts out making angle θ with $|k_o\rangle^\perp$; and each application of $-WV_{\rm in}$ increases that angle by 2θ. So, if we apply $-WV_{\rm in}$ to $|\phi\rangle$ a total of s times, the resulting vector will make angle $(2s+1)\theta$ with $|k_o\rangle^\perp$. Now apply the operator $-WV_{\rm in}$ to $|\phi\rangle$ a total of s times, where s is that number such that $(2s+1)\theta$ is closest to $\pi/2$. Then this number of times will satisfy $s \leq \pi/(4\theta) \leq (\pi/4)\sqrt{N}$, where in the second inequality we used $\theta \geq \sin\theta = \frac{1}{\sqrt{N}}$. Having applied $-WV_{\rm in}$ to $|\phi\rangle$ this many times, the resulting vector in this plane will be within angle θ of $|k_o\rangle$. Let us now make an observation on this final vector, via the $|k\rangle$-basis for $H_{\rm in}$. The probability that this observation results in k_o, by what we just observed, is $\geq \cos^2\theta = 1 - \frac{1}{N}$. That is, our chances are excellent that this single observation on $H_{\rm in}$ will find the needle.

So, to summarize, if we apply, to initial state $|\phi\rangle$ $\frac{1}{\sqrt{2}}\{|{\rm no}\rangle - |{\rm yes}\rangle\}$ in $H_{\rm in} \otimes H_{\rm out}$, the operator $-WV$ a number of times not exceeding $\frac{\pi}{4}\sqrt{N}$, and then observe the resulting state via the $|k\rangle$-basis, we will, with probability at least $1 - \frac{1}{N}$ (i.e., almost certainly), obtain the needle, k_o. Note that we only have to run the computer (i.e., apply V) a number of times proportional to $\sqrt{N}$ — *not* to N itself. It does indeed appear that there has been a significant gain in efficiency. This is an example of a quantum-assisted computation.

Note that if you are impatient — insisting on making $|k\rangle$-basis observations between the computer runs ("just to see how things are going"), then you will destroy this effect. This is similar to the familiar "watched pot never boils" parable in quantum mechanics.

16 Grover Construction: Six Issues

In the previous section, we gave an example of a construction that appears to show quantum mechanics providing a clear gain in efficiency over a non-quantum computation. We here discuss six issues pertaining to that construction.

16.1 Initial State

The construction requires that the registers be placed, initially, in state $|\phi\rangle \ \frac{1}{\sqrt{2}}\{|\text{no}\rangle - |\text{yes}\rangle\}$. Is it feasible to build this state?

The state of H_{out} would not seem to be much of a problem: After all, this is merely a 2-dimensional Hilbert space. So, for example, we could represent this space physically as the spin-states of a spin-1/2 particle, designating $|\text{no}\rangle$ and $|\text{yes}\rangle$ as the states corresponding to the spin aligned or anti-aligned in a given direction. Then $\frac{1}{\sqrt{2}}\ \{|\text{no}\rangle - |\text{yes}\rangle\}$ would be the state in which the spin is aligned in a certain orthogonal direction.

But the state $|\phi\rangle \in H_{\text{in}}$ is more complicated. After all, this is a superposition of N states. To construct these states one at a time, and then "superpose them" (whatever that means) is a job that threatens to have difficulty N, i.e., to overwhelm the difficulty of running the computer $\sqrt{N}$ times. Here is a device — common in this subject — to circumvent this problem. Fix, once and for all, a 2-dimensional Hilbert space, with basis $|0\rangle, |1\rangle$ (not to be confused with the vectors of the same name in H_{in}). So, e.g., this H might be the spin-states of a spin-1/2 system. Let $N = 2^n$ for some positive integer n. [To achieve this — at most a doubling of the number of input states — should not cause too much additional complication.] Now set

$$H_{\text{in}} = H \otimes H \otimes \cdots \otimes H, \tag{7}$$

where a total of n copies of H appear on the right[8]. Note that this gives the correct dimension for H_{in}. Now consider a typical state, e.g., $|0\rangle|1\rangle|1\rangle|0\rangle \cdots |0\rangle|1\rangle$ (total of n factors) in the Hilbert space on the right. We identify this with the state $|k\rangle$ of H_{in}, where $k = 0110\cdots01$ in base 2. The k's that result in this way range from 0 (for $00\cdots0$) to $2^n - 1$ (for $11\cdots1$); and so we indeed obtain in this way the basis we want for H_{in}. Under this identification, the construction of the state $|\phi\rangle \in H_{\text{in}}$ is quite easy: It is a simple product

$$|\phi\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\ \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\ \cdots\ \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$$

[8]Note that we could not, e.g., let the H's be simply the spin-states for n identical spin-1/2 particles, because the states on the right in (**??**) are not in general antisymmetric under particle-interchange. However, we could, e.g., have a system of n electrons occupying n energy levels (say, in an atom); where each H, referring some energy level, gives the spin-state of the occupant of that level.

of the states $\frac{1}{\sqrt{2}}\,(|0\rangle + |1\rangle)$ for each of the H-factors. This follows, expanding the right side, and using the definitions of $|k\rangle$ and $|\phi\rangle$. Thus, with this choice of how H_{in} is to be structured, the construction of the state $|\phi\rangle$ should be relatively easy. We note that this construction could have been carried out[9] with any fixed dimension for the factor-Hilbert spaces H.

16.2 Final Observation on H_{in}

The construction requires that, at the end of the computer-runs, an observation be made on H_{in} via the $|k\rangle$-basis. Is it feasible to make such an observation?

Yes, it is. Denote by A the Hermitian operator $|0\rangle 0\langle 0| + |1\rangle 1\langle 1|$ on H (so observation of A is observation of H via its natural basis). For example, if H is spin-states, and the basis is spin-component in a certain direction, then A would be the observation of spin-component in that direction, plus $1/2$. Now consider the following Hermitian operator on $H \otimes \cdots \otimes H$: Operator $(2^{n-1}A)$ applied to the first H-factor, plus operator $(2^{n-2}A)$ applied to the second H-factor, and so on, until reaching finally operator $(2^0 A)$ applied to the last H-factor. [Here, we are regarding these operators on the H-factors as operators on the tensor product in the manner described in Sect. 14.] The resulting sum can, via (**??**), be regarded as an operator on H_{in}; and we note that it does indeed have the $|k\rangle$ as its eigenstates. [In physical terms, observe the first H-component and multiply by 2^{n-1}; the second, by 2^{n-2}; etc., and add. The result will be precisely the k-value of that state.] We would expect to have no difficulty in making an observation of H via this operator A; and, therefore, no difficulty in making an observation of H_{in}, so constructed, via its $|k\rangle$-basis.

16.3 Building the Operator W

The construction requires that we apply unitary operator $W = I - 2|\phi\rangle\langle\phi|$ to H_{in}. Is it feasible to build and apply such an operator?

[9] In fact, the different H's in the product could, if we so desired, be assigned different dimensions. **Exercise.** Set up a system of arithmetic in which each of the various digits of an integer refers to different number-system base. Figure out how to add in this system (which turns out to be quite simple!).

Note that this by no means follows immediately from the prior point: The mere fact that we feel capable of placing H_{in} in state $|\phi\rangle$ does not lead directly to an interaction on H_{in} that shifts each state $|\psi\rangle \in H_{\text{in}}$ to state $W|\psi\rangle \in H_{\text{in}}$. In order to build the operator W, we proceed as follows.

We first require a few preliminaries. We introduce a more convenient basis for H: $|\alpha\rangle = \frac{1}{\sqrt{2}}(|1\rangle + |0\rangle)$, $|\beta\rangle = \frac{1}{\sqrt{2}}(|1\rangle - |0\rangle)$. In terms of this basis, we have $W = I - 2|\alpha\rangle \cdots |\alpha\rangle\langle\alpha| \cdots \langle\alpha|$. Next, we introduce a unitary operator T on $H \otimes H \otimes H$, with the following action: $T(|\alpha\rangle|\alpha\rangle|\alpha\rangle) = |\alpha\rangle|\alpha\rangle|\beta\rangle$, $T(|\alpha\rangle|\alpha\rangle|\beta\rangle) = |\alpha\rangle|\alpha\rangle|\alpha\rangle$, while T the identity on the other six basis elements of $H \otimes H \otimes H$. That is, this operator T, which is called the *Toffoli gate*, flips the third H-state if and only if the first two H-states are[10] both $|\alpha\rangle$; and so, e.g., we have $T^2 = I$. Let us denote by $H_1, H_2, \cdots, H_n$ the n H's in the tensor product that is H_{in}. We now introduce a second Hilbert space, $H_{\text{scratch}} = \tilde{H}_3 \otimes \tilde{H}_4 \otimes \cdots \otimes \tilde{H}_{n+1}$, where each of the $\tilde{H}$ in this tensor product is also a copy of our basic Hilbert space H. This is the Hilbert space in which we shall carry out scratch work. Thus, our full Hilbert space is now $H_{\text{in}} \otimes H_{\text{scratch}}$, a tensor product of $2n-1$ copies of H. Now consider the following operator on this tensor product:

$$\mathcal{W} = T(H_n, \tilde{H}_n, \tilde{H}_{n+1})T(H_{n-1}, \tilde{H}_{n-1}, \tilde{H}_n) \cdots T(H_3, \tilde{H}_3, \tilde{H}_4)$$
$$\times T(H_1, H_2, \tilde{H}_3). \tag{8}$$

We note that this operator, as a composition of unitary operators, is unitary. Let us now begin with an arbitrary state in H_{in}, but with H_{scratch} in the state $|\tau\rangle = |\beta\rangle|\beta\rangle \cdots |\beta\rangle \frac{1}{\sqrt{2}}(|\alpha\rangle - |\beta\rangle)$. Let us apply to this state the operator (**??**), and see what happens. The rightmost operator T in this composition will place $\tilde{H}_3$ (which began in state $|\beta\rangle$) in state $|\alpha\rangle$ if and only if the H_1- and H_2-states are both $|\alpha\rangle$. The next T, reading from right to left, will place $\tilde{H}_4$ in state $|\alpha\rangle$ if and only if H_3 and $\tilde{H}_3$ are both in state $|\alpha\rangle$, i.e., if and only if H_1, H_2, and H_3 are all in state $|\alpha\rangle$. And, similarly, the next T will place $\tilde{H}_5$ in state $|\alpha\rangle$ if and only if all four of H_1, H_2, H_3 and H_4 are in state $|\alpha\rangle$.

[10]The general state in $H \otimes H \otimes H$ is, of course, not one in which each of the H's is in a particular state ($|\alpha\rangle$ or $|\beta\rangle$), but rather is a superposition of all eight possible combinations of individual H-states. We often describe operators, such as this T, by giving their actions on each of the combinations that appear in this superposition. Thus, when we say, e.g., "the first two H-states are ...", we really mean "that term, in the superposition, in which the first two H-states are ..."

Continue in this way, working from right to left in (**??**). Recall that the last $\tilde{H}$, $\tilde{H}_{n+1}$ begins in state $\frac{1}{\sqrt{2}}(|\alpha\rangle - |\beta\rangle)$ rather than $|\alpha\rangle$. Thus, in the last step, an attempt to "flip" the $\tilde{H}_{n+1}$-state will merely introduce a minus sign. We conclude: The operator $\mathcal{W}$ of (**??**), acting on a state $|\psi\rangle|\tau\rangle \in H_{\text{in}} \otimes H_{\text{scratch}}$, where $|\psi\rangle$ is any state in H_{in}, and $|\tau\rangle$ is the state in H_{scratch} given above, indeed generates a sign change if all the H's are in state $|\alpha\rangle$, and no sign change otherwise.

The operator $\mathcal{W}$, so constructed, is our candidate for W. Of course, it acts, not merely on the Hilbert space H_{in} (as W does), but rather on $H_{\text{in}} \otimes H_{\text{scratch}}$ Nevertheless, it does seem to have the right action and so, it appears, would seem to suffice for the Grover construction.

But this appearance is misleading: The above candidate, $\mathcal{W}$, will not work as a proxy for W, for the following reason. Some scratch work for this calculation was left in the auxiliary Hilbert space H_{scratch}. That is, the final state, after application of $\mathcal{W}$ is an entanglement of H_{in}-states and H_{scratch}-states. Consider, for example, $n = 4$. Then if the initial state of H_{in} was $|\alpha\rangle|\alpha\rangle|\beta\rangle|\alpha\rangle$, say, then the final state of H_{scratch} will be $|\alpha\rangle|\beta\rangle\ \frac{1}{\sqrt{2}}(|\alpha\rangle - |\beta\rangle)$; while if the initial state of of H_{in} was $|\alpha\rangle|\beta\rangle|\beta\rangle|\alpha\rangle$, then the final state of H_{scratch} will be $|\beta\rangle|\beta\rangle\ \frac{1}{\sqrt{2}}(|\alpha\rangle - |\beta\rangle)$. This entanglement, we claim, will destroy the working of the Grover construction. To see this, consider Eqn. (**??**), which gives the result of the first application of $-WV_{\text{in}}$ to $|\phi\rangle$: a rotation $|\phi\rangle$ through angle 2θ. The key to the construction is that the next application of $-WV_{\text{in}}$ (as well as each successive application) must rotate through an additional angle 2θ. But, in order for this to happen, there must occur cancellation between the $|\phi\rangle$'s and $|k_o\rangle$'s that arise from application of $-WV_{\text{in}}$ to the two terms on the right in (**??**). Now consider what happens if the W on the left in (**??**) is replaced by $\mathcal{W}$. Then the terms on the right side of this equation will become entangled with various elements of H_{scratch}. Therefore, on the next application of $-WV_{\text{in}}$ the necessary cancellations on the right will not take place. The Grover construction will thus fail.

In order to obtain an effective W, we proceed as follows. Set

$$\mathcal{W}' = T(H_3, \tilde{H}_3, \tilde{H}_4) \cdots T(H_{n-1}, \tilde{H}_{n-1}, \tilde{H}_n)\ \mathcal{W}. \tag{9}$$

That is, $\mathcal{W}'$ first applies $\mathcal{W}$, and then applies all the operators of $\mathcal{W}$, save the leftmost, in reverse order. It is easy to check that this procedure undoes the entanglement. That is, we have $\mathcal{W}'(|\psi\rangle|\tau\rangle) = |W\psi\rangle|\tau\rangle$ for any $|\psi\rangle \in H_{\text{in}}$,

where $|\tau\rangle \in H_{\text{scratch}}$ is the initial state given above. *This* $\mathcal{W}'$, then, can be used in place of W in the Grover construction.

We conclude, then, that the operator W on H_{in} in the Grover construction can indeed be built, by introducing an auxiliary Hilbert space H_{scratch}, and applying the Toffoli gate (an operator on $H \otimes H \otimes H$) a total of $2n-3$ times. So, it would seem that the operator W is feasible — provided the operator T is feasible. We shall return to this last issue shortly.

16.4 Building the Operator V

No real computer, it might be argued, operates by applying some unitary operator V to $H_{\text{in}} \otimes H_{\text{out}}$, as in the Grover construction. After all, real computers use irreversible operations (such as placing bits in locations). How, then, are we to construct and interpret the operator V?

Here is a model for how a computer might operate. We introduce an additional Hilbert space, H_{com}, to represent the computer states. Then the total Hilbert space is $H_{\text{in}} \otimes H_{\text{out}} \otimes H_{\text{comp}}$. The running of the computer will then be represented by some unitary operator, $\mathcal{V}$, on this Hilbert space. Let us fix a vector $|\psi_{\text{init}}\rangle \in H_{\text{comp}}$, to represent the initial state of the computer. Then, in the Grover case (i.e., with H_{in} spanned by $|0\rangle \cdots |N-1\rangle$ and H_{out} by $|\text{no}\rangle, |\text{yes}\rangle$)), the action of a suitable $\mathcal{V}$ would be as follows:

$$\mathcal{V}(|k\rangle|\text{no/yes}\rangle|\psi_{\text{init}}\rangle) = |k\rangle|\text{no/yes}\rangle|\psi_k\rangle, \tag{10}$$

where the H_{out}-state on the right is $|\text{no}\rangle$ or $|\text{yes}\rangle$ depending on whether the H_{out}-state on the left is $|\text{no}\rangle$ or $|\text{yes}\rangle$, and also on whether or not $k = k_o$. The $|\psi_k\rangle \in H_{\text{comp}}$ on the right in (**??**) is the final state in which the computer finds itself, depending on the k-value on the left (and also on the choice of initial state in H_{out}, which we suppress). This operator $\mathcal{V}$ is unitary, and so invertible. Thus, we are suggesting, the operation of any computer must always be reversible. [Indeed, in a world governed by quantum mechanics, this is necessary, for dynamics therein is described by an (invertible) unitary operator.] Things don't appear to be this way in practice only because we fail to take into account how large and complicated H_{comp} can be. It includes not only the states of the chips, wires, fan, etc within the box, but also (if, say, the computer is plugged in) the states of the electric company, and then of its employees, etc. By the time all this dust settles, things look pretty irreversible.

Unfortunately, the operator $\mathcal{V}$ of (**??**) will not serve as a proxy for the operator V of the Grover construction. The problem is that $\mathcal{V}$ introduces entanglements between $H_{\rm in} \otimes H_{\rm out}$ on the one hand and $H_{\rm comp}$ on the other, as reflected in the dependence of the final computer state, $|\psi_k\rangle$, in (**??**) on k. These entanglements, in the same manner as for $\mathcal{W}$ in the discussion above, will interfere with the cancellation that must take place in Eqn. (**??**), and will thereby cause the Grover construction to fail. In order to avoid these entanglements, we must, e.g., so design our computer that the final computer state, say $|\psi_{\rm final}\rangle$, is independent of $|k\rangle$. Then, when it comes time to repeat the computation, we could either apply some special treatment to the computer to restore its initial state to $|\psi_{\rm init}\rangle$, or discard that computer entirely, bringing in another with the initial state $|\psi_{\rm init}\rangle$ already preinstalled[11]. Best would be if we could arrange that $\mathcal{V}$ automatically, at the end of each run, returns the computer to state $|\psi_{\rm init}\rangle$, ready for the next run.

So, in any case, in order to carry out the computation implicit in the Grover construction, we shall have to produce a computer that does not introduce entanglements between computer states and in-out states. This is definitely *not* the computer on your desk! We shall have to build our computer anew. The danger we face is that the building and operating of such computers consumes resources — in particular, time — and we must be careful that this consumption does not overwhelm the apparent savings we derive from using quantum mechanics.

Recall that the Hilbert space $H_{\rm in} = H \otimes \cdots \otimes H$, in the Grover construction, has large dimension, 2^n. Our computer must interact with this large Hilbert space, but do so relatively efficiently. It would be of great help if we could design our computer to interact, not with all n of the H's at once, but rather with only a few at a time. Does this restriction entail a restriction on the possible unitary operators we can generate on $H_{\rm in} \otimes H_{\rm out}$? The following shows that it does not.

Theorem. Let H be a finite-dimensional Hilbert space. Then any unitary

[11] Why not get rid of these awkward entanglements, not by searching for a clever $\mathcal{V}$, but rather by simply discarding the computer after each run, bringing in a new computer, with $|\psi_{\rm init}\rangle$ preinstalled, for the next run? The problem with this maneuver is that the act of discarding a system entangled with another places the latter system in a mixed state, as described by a density operator. But a mixed state for $H_{\rm in}$ destroys the cancellation, and so the Grover construction, as surely as does entanglement

operator on $H \otimes H \otimes \cdots \otimes H$ is equal to a product of unitary operators, each of which acts on at most two of the H-factors in this tensor product.

Of course, different combinations of the two H-factors are allowed for the different unitary operators in this product. Our proof of the theorem will make use of three facts.

Lemma 1. Every Hermitian operator on $H \otimes \cdots \otimes H$ is a linear combination of operators of the form $A \otimes \cdots \otimes B$, where $A, \cdots, B$ are Hermitian operators on H.
Lemma 2. Every Hermitian operator on a Hilbert space is a linear combination of commutators of Hermitian operators, and the identity I.
Lemma 3. Fix a connected Lie group G, and a collection of one-parameter subgroups of G. If the generators of these subgroups generate the entire Lie algebra of G, then the subgroups themselves generate the entirety of G.

Lemma 1 is easy to prove by a dimensional argument, using that the dimension of the (real) vector space of Hermitian operators on a Hilbert space is equal to the square of the dimension of that (complex) space. Let there be n H's in the tensor product, each of dimension m. Then the Hilbert space $H \otimes \cdots \otimes H$ has dimension m^n, and so the vector space of Hermitian operators on this space has dimension $(n^m)^2$. The Hermitian operators of the form given in the Lemma form a subspace of this space, and it has dimension (dimension of Hermitian operators on $H)^m = (n^2)^m$. These dimensions are equal, and so the subspace is the entire vector space[12]. Lemma 2 (which, apparently, has little independent interest) follows by direct construction. For $m = 2$, for example, it is the statement that any linear combination of spin-operators is a commutator of two such linear combinations. In Lemma 3, the Lie algebra of a Lie group is the tangent space at its identity element. The "generator" of a one-parameter subgroup is that element of the Lie algebra given by the tangent to that curve at the identity. The Lie algebra "generated" by these generators is the collection of all elements that can be obtained by using linear combinations and brackets on the generators

[12]Lemma 1 is not quite as empty as it may appear at first sight. To see this, you might try to write the Hermitian operator that switches the two H-states in $H \otimes H$ in the form guaranteed by the Lemma.

of the one-parameter subgroups. And, finally, that the elements of these subgroups "generate" the entirety of G means that every element of G can be written as a (finite) product of such subgroup-elements. This Lemma, in other words, states that if you can get the entire group from the subgroups "infinitesimally close to the identity", then you can indeed get the entire group from the subgroups "everywhere"[13]. [This is the sort of thing that would normally be used, without mention, in a physics course.]

The theorem is very easy to prove from the three Lemmas. Consider, say, $n = 3$. We have, for A, B, C, and D any Hermitian operators on H, and a any real number

$$[A \otimes I \otimes C,\ I \otimes B \otimes D] + a\ A \otimes B \otimes I = A \otimes B \otimes ([C, D] + aI). \qquad (11)$$

where "[,]" denotes i times the commutator. Each of the operators that appears on the left contains an "I", and so is a generator of unitary operators on $H \otimes H \otimes H$ that act on only two factors. By Lemma 2, the right side of (**??**) includes the general tensor product of three Hermitian operators on $H \otimes H \otimes H$; and, by Lemma 1, these span all Hermitian operators on the tensor product. The result (for $n = 3$) now follows from Lemma 3. The case of general n is by induction on n, repeating the construction of (**??**) at each step.

It seems likely that the product (whose existence is guaranteed by the Theorem) involves no more than 5^{n-2} factors (perhaps substantially fewer), by an argument that traces the mechanism of Lemma 3. Unfortunately, this number grows quickly with n. The Theorem is also true (suitably modifying Lemma 1) when the H's in the tensor product have different dimensions.

So, we may expect to build our computer out of operators that act on H-factors two at a time. But is it feasible to construct even these operators? Let us take, as an example, the case in which H is 2-dimensional, representing the spin-states of an electron. Then the general Hermitian operator on H is $\vec{s} \cdot \vec{\sigma} + b\ I$, where $\vec{s}$ is any vector in 3-space, $\vec{\sigma}$ is the vector (Pauli) spin operator, and b is any real number. [Note that these form a 4-dimensional

[13] As an example, let G be the rotation group, let one one-parameter subgroup be the rotations about some vector $\vec{s}$, and let the another be rotations about some other, independent, vector $\vec{t}$. Then, since the Lie bracket of the corresponding infinitesimal rotations is simply an infinitesimal rotation about $\vec{s} \times \vec{t}$, the hypothesis of Lemma 3 is satisfied. The Lemma asserts in this case that *every* rotation can be written as some product of various rotations about $\vec{s}$ and $\vec{t}$. This fact is the basis of Euler angles.

vector space, as required.] The Hermitian operator $\vec{s} \cdot \vec{\sigma}$ generates the family of unitary operators, written $e^{i\ \vec{s} \cdot \vec{\sigma}}$, that correspond to rotations in space about the vector $\vec{s}$ as axis; while $b\ I$ generates the family, written e^{ib}, that correspond to overall phase-changes (which have no physical significance).

The unitary operators on a single H can be constructed physically as follows. The unitary operators corresponding to rotations about $\vec{s}$ result from applying to the electron a magnetic field in the $\vec{s}$-direction, for such a field causes the electron, by virtue of its angular momentum and magnetic moment, to precess about the magnetic-field direction. The product of the field-strength and the time for which the interaction is turned on determine the magnitude of this rotation.

But, in order to invoke the Theorem, we must also construct the unitary operators on $H \otimes H$, i.e., on the two-electron system. It should be clear that merely subjecting the two electrons, each to its own magnetic field, will not suffice. We must introduce some sort of direct interaction between the two electrons. One such is what is called the spin-spin interaction. The corresponding Hermitian operator on $H \otimes H$ is $\vec{\sigma_1} \cdot \vec{\sigma_2}$ where $\vec{\sigma_1}$ and $\vec{\sigma_2}$ denote the spin operator acting on the first and second factor in $H \otimes H$, respectively. [Strictly speaking, we should include a "$\otimes$" between the two σ's in this expression; but the dot gets in the way.] This particular interaction actually occurs in nature: If the two electrons are merely brought close together[14], then, by virtue of the electromagnetic interaction between their magnetic moments, the electrons interact in just the manner we have described. The corresponding unitary operator may be written $e^{ia\ \vec{\sigma_1} \cdot \vec{\sigma_2}}$, where the number a is determined by how close together the electrons are placed, and for how long.

That these two physical operations — placing one or both electrons in a magnetic field, and allowing the electrons to interact electromagnetically — suffice to generate all possible two-electron interactions now follows from:

Theorem. Let H be a 2-dimensional Hilbert space. Then every unitary operator on $H \otimes H$ is equal to some (finite) product of the operators e^{ib}, $I \otimes e^{i\ \vec{s} \cdot \vec{\sigma_1}}$, $I \otimes e^{i\ \vec{s} \cdot \vec{\sigma_2}}$, and $e^{ia\ \vec{\sigma_1} \cdot \vec{\sigma_2}}$, where $\vec{s}$ is any vector in 3-space and a and

[14]This could be done, for example, by keeping the electrons in boxes, with H representing the spin-state of the occupant of a given box; and then moving the boxes into close proximity.

b are any real numbers.

The proof is virtually identically to that of the earlier Theorem, using the Lemmas in the same way. In this case, Eqn. (**??**) is replaced by

$$-[\vec{t}\cdot\vec{\sigma_1}\otimes I,\ [\vec{s}\cdot\vec{\sigma_1}\otimes I,\ \vec{\sigma_1}\cdot\vec{\sigma_2}]] + (\vec{s}\cdot\vec{t})\ \vec{\sigma_1}\cdot\vec{\sigma_2} = (\vec{s}\cdot\vec{\sigma_1})\otimes(\vec{t}\cdot\vec{\sigma_2}), \quad (12)$$

where we have used the fact that $[\vec{s}\cdot\vec{\sigma},\ \vec{t}\cdot\vec{\sigma}] = i(\vec{s}\times\vec{t})\cdot\vec{\sigma}$. Taking linear combinations involving the right side of (**??**) and the Hermitian operators $\vec{s}\cdot\vec{\sigma_1}\otimes I$, $I\otimes\vec{s}\cdot\vec{\sigma_2}$, and $I\otimes I$ reproduce the entire Lie algebra of $H\otimes H$; which is just what we need to complete the proof.

Here are a couple of examples. Consider the operator W on $H_{\rm in} = H\otimes\cdots\otimes H$, where H is taken as the two-dimensional Hilbert space of spin-1/2 states. It follows, from the two theorems above, that this W is equal to a composition of our basic unitary operators: That (on a single H) generated by a magnetic field, and that (on two H's) generated by the spin-spin interaction. It seems likely that the number of such basic operators that must be composed to construct W in this manner increases exponentially in n. Note that this construction of W is different from that of Sect. 16.3, for there we made use of an auxiliary Hilbert space $H_{\rm scratch}$, whereas here there is none. Next, note, again from the two theorems above, that the Toffoli operator, T, on $H\otimes H\otimes H$ is also equal to a product of the basic operators on H. Having constructed T from the basic operators, we may then proceed to construct W from the basic operators, using the strategy of Sect. 16.3. While this alternative construction of W requires an auxiliary Hilbert space $H_{\rm scratch}$, it does have the advantage that the number of basic operators required grows only linearly with n. We may, in addition to W, also construct $V_{\rm in}$, in the following manner. Denote by U the unitary operator on H with action $U|0\rangle = |1\rangle, U|1\rangle = |0\rangle$. Suppose that the needle, written in base 2, is, say, $k_o = 01001\cdots01$. Then set $\mathcal{U} = U\otimes I\otimes U\otimes U\otimes I\cdots\otimes U\otimes I$, where the U's and I's on the right correspond to the digits in this expression for k_o. We then claim that $V_{\rm in} = \mathcal{U}W\mathcal{U}$. This is easy to check: $\mathcal{U}$ sends $|k_o\rangle$ to $|1\rangle|1\rangle\cdots|1\rangle$ (but other $|k\rangle$'s to something else); and then W produces a minus sign (but, for other $|k\rangle$'s, a plus sign); and then the final $\mathcal{U}$ restores the original state. Note that, in this construction of $V_{\rm in}$, the number of operators that must be composed grows only linearly with n.

The discussion above shows that there will in general be a variety of ways to construct a given unitary operator out of some set of basic operators. Some

ways may involve an auxiliary Hilbert space (in which case we must avoid entanglement) and some not; some may involve the composition of a large number of basic operators and some a smaller number. But, unfortunately, none of this is what we really want for the Grover construction. The V_{in} above, for example, is completely useless for our purposes, because it requires that you already "know" k_o, and this is exactly what you are not supposed to know. What we need is, not a variety of ways to construct some "given" unitary operator on a Hilbert space out of basic operators, but rather a way to convert programs into operators. In the Grover case, for example, we begin with a computer program that checks whether or not a given k is the needle; and we wish to convert that program to a suitable unitary operator V_{in}. Here is a general summary of what we are looking for.

> Let there be given a program P that accepts as input nonnegative integers, and returns nonnegative integers. We may register the inputs and outputs in Hilbert spaces H_{in} and H_{out}, each of which is a (finite) tensor product of some finite-dimensional Hilbert space H with itself. We wish to convert the program P into a unitary operator on $H_{\text{in}} \otimes H_{\text{out}}$, such that this operator "computes" the output from the input in the manner of P, and does so with substantially the same difficulty function as that of P. This unitary operator may require an auxiliary Hilbert space H_{scratch}, but if it does then the operator must be such that the $H_{\text{in}} \otimes H_{\text{out}}$-part of the final state is not entangled with the H_{scratch}-part.

It is not clear how to make this summary into a precise statement. What, for example, does "... in the manner of ..." mean; and what is the "difficulty function" of a unitary operator? We have in mind some sort of compiler, which turns command-lines in the program into compositions of some basic unitary operators on the Hilbert spaces. But it is not clear how this is to work. What, for example, are the operator-equivalents of APPEND and DELETE? Even more difficult would be to find an operator-equivalent of IF (LAST $C(S) == x$) THEN SKIP n LINES. In any case, having constructed such a compiler, then we might be able to define a suitable difficulty function in terms of the number of basic operators in the composition. And even after all this, we would have to contend with the fact that *programs* accept as input

arbitrary integers, while our *operators* act on just finite tensor products. All this is somewhat reminiscent of the issue, discussed in Sect. 13, of compiling programs in machine language. It might be worthwhile to try to resolve that issue first, as a prerequisite to this one.

There seems to be a sense, in this field, that there may exist some sort of construction along the lines outlined above.

16.5 Errors

Errors abound in the Grover construction. They can appear in the setting up of the initial state, in the application of the operators W and V, and in the observation on the final state. Errors could arise, for example, from imperfections in the apparatus; or from quantum tunneling causing interactions between the H-states and the thermal fluctuations in the outside world. How shall we take such errors into account?

Of course, errors abound everywhere in physics. But here, because of the kinds of questions we are asking, this issue seems particular compelling. Consider a computation, and suppose that, on the initial run, the input string S is such that we require just 10 steps. Then we can afford, for this run, to be relatively cavalier about errors. But the next run, for another S, might involve 1,000 steps, requiring, in order to keep the effects of errors in check, that we purchase new and better equipment. And still another run might involve 1,000,000,000 steps, requiring that we cool the entire Earth down to 0.03^o K and move the Sun over to another part of the Galaxy. All of these extra precautions take time and effort, and so should be taken into account in the difficulty function. But how will we be able ever to include such things? How, for example, does the difficulty of these various precautions scale with the number of steps in the computation? Similar issues arise already in ordinary computing: Bits are sometimes recorded incorrectly; and the longer the calculation the greater the care that must be exercised in this regard.

We shall simply ignore the effects of errors, not out of any conviction these effects are likely to be unimportant, but rather because we do not know how to do anything else.

16.6 What Is The Problem?

The Grover construction, as you have undoubtedly noticed, does not compute any problem at all, at least, not as we have defined that term. A "problem" entails an output for *every* possible input string; while the Grover construction searches for the needle in a finite haystack. Finite input sets are not very interesting: All problems based on them are computable, and all difficulty functions on them are equivalent.

The obvious way to respond to this situation would be to modify the construction to apply to a variety of haystacks. Given N, we first construct the Hilbert space H_{in} (as a tensor product of about $\log_2 N$ copies of a 2-dimensional H), then build our computer (i.e, construct our operator V). We are now prepared to apply the Grover construction to find the needle. Of course, the building of the computer (i.e., of the operator V) is an additional burden, which would, presumably, be included in the difficulty function for this computation.

So let us suppose, then, that we have suitably modified the Grover construction along these lines. We would then be in a position to ask the key question: Is there any problem for which the Grover construction, so configured, is more efficient than any computation-method not using quantum mechanics? Here is a precise mathematical assertion that reflects these ideas.

Assertion. Let π be any problem that accepts as input a pair (N, k), where N is a positive integer and k is one of the integers $(0, 1, \cdots, N-1)$, and returns either "yes" or "no", such that: For each N, there is one and only one of those k's (call it k_o) for which $\pi(N, k)$ is "yes". Let π' be the problem that accepts as input any positive integer N, and returns this k_o. Let P be any program (written, say, in the language of Sect. 12) that computes π, with difficulty function f; and set $h(N) = \max_k f(N, k)$. Then there exists a program P' that computes π', with difficulty function f' satisfying $f' \leq \sqrt{N}h$ (in the sense of difficulty functions).

The idea of this assertion is the following. Think of the problem π as a sequence of needle-in-the-haystack challenges. Here, N represents the haystack itself, and k is a needle-candidate for that haystack. Thus, the problem π takes such a haystack and needle candidate, and returns the answer to the following question: "Is that k the needle for that haystack?" The program

P computes this π. That is, P tests whether, for a given haystack, a given candidate for the needle is indeed the correct needle, (k_o). By contrast, the problem π' simply announces, given the haystack, what the needle is for that haystack. And the program P' computes that problem π'. That is, P' takes the haystack and finds the needle.

Note that, given a program P that computes π, we can immediately write a program P' that computes π': For each given N, this P' merely runs P on the pair (N, k), for each $k = (0, 1, \cdots, N-1)$ in turn. It then finds that k for which P returns "yes", and announces that k-value. [Thus, in particular, if π is computable, then automatically π' is.] What will be the difficulty function, f', of this program P'? Fix N. Then the maximum difficulty to check a single k-value is the $h(N)$ given in the assertion. Since P' must run P at most N times, we have $f'(N) \leq Nh(N)$. We conclude: If, in the inequality of the assertion above, "$\sqrt{N}$" were replaced by "N", then the assertion would be true, choosing for P' this program that simply runs P a total of N times.

But the assertion as it stands, says that there always exists a shortcut P' that is *better*, in a suitable sense, than the naive P' constructed above. It says that you can always discover some way of finding the needle with difficulty not exceeding the maximum P-difficulty to check one candidate, multiplied by the *square root* of the number of candidates that P' would have to check. The assertion asserts, in other words, that given any family of needle-challenges, and a way to meet those challenges by trial-and-error, then there exists a way to meet those challenges that is more efficient than trial-and-error, by a factor of $\sqrt{N}$.

Why this $\sqrt{N}$? It comes from the Grover construction! Imagine that we had somehow come up with a counterexample to the assertion above. That is, we have a problem π, of the type indicated above, and program P that computes it, such that there does not exist any shortcut program P', in the sense of the assertion. Fix N, and consider the action of P on (N, k), for that N. Let us next imagine that we were able to simulate this action of P by a suitable unitary operator, V, on $H \otimes \cdots \otimes H$ (n times, where $2^n \geq N$), and that the total "difficulty" required to apply this operator was just $h(N)$, i.e., the maximal difficulty that the ordinary program P encounters for the various (N, k), with this fixed N. And finally, let us suppose further that the additional difficulty of building the quantum system is sufficiently small that it may be disregarded. Then the Grover construction would find the needle (with very large probability) with a total difficulty of $\sqrt{N}h(N)$ (since, as we

saw in Sect. 15, the computer would have to be run only $\sqrt{N}$ times). But we began with the assumption that this P is a counterexample to the assertion, i.e., that it is is such that exists no shortcut P' with $f'(N) \leq \sqrt{N}h(N)$. What this means, in other words, is that there is no regular program that solves this problem more efficiently than does the Grover construction.

We conclude: A counterexample to the assertion above would provide a road map for finding (via Grover) an example in which a quantum-assisted computation is more efficient than any computation of the same problem without a quantum-assist. If the assertion were true, on the other hand, then this result would considerably diminish the prospects for using the Grover construction in this way.

We emphasize that the assertion above is a statement in mathematics: It does not involve quantum mechanics, nor any details of how computations work. It is either true or false. I have neither a proof nor a counterexample to this assertion. Below are three examples of (failed) attempts to construct a counterexample, which are intended to give a sense of how the assertion works.

For the first example, let $\pi(N, k)$ be "yes" if and only if $k = N/2$ (or $(N+1)/2$, if N is odd). Thus, program P would, on receiving (N, k), multiply k by 2, and see if the result is N (or $N+1$). The difficulty function is $\log(N)$ (i.e., effectively, the number of digits of N, independent of k); and so we have $h(N) = \log(N)$. This program P is not a counterexample. Let P' be the program that accepts positive integer N, simply computes $N/2$ (or $(N+1)/$, and returns that integer. The difficulty function for this P' is also $f'(N) = \log(N)$, and so we certainly have $f'(N) \leq \sqrt{N}h(N)$. This candidate for a counterexample was hopeless right from the beginning. Any time the structure of P is "do some computation involving N, and then check to see if the result matches k", then you will never end up with a counterexample. Program P' will overhear this strategy and proceed to compute the needle k_o directly from N in the same manner, ending up with difficulty function given by $f'(N) = h(N)$, and thus satisfying the inequality of the assertion.

For the second example, let $\pi(N, k)$ be "yes" if and only if k is the largest prime $\leq (N-1)$. Thus, program P would, on receiving (N, k), first check to see if k is prime (reporting "no" if it is not), then check the integers from $k+1$ to to $N-1$ for primeness (reporting "no" if any is prime), and otherwise reporting "yes". The difficulty function, $f(N, k)$, of this program has complicated k-dependence. But in any case, denote by $h(N)$ the greatest

difficulty encountered as k ranges from 0 to $(N-1)$. This program P is not a counterexample. Let P' be the program that works downward from $N-1$, checking each integer for primeness, and reporting the first prime it finds. This P', then, computes the problem π'. The number of steps required by P' will be the same as for P to check candidate k_o, i.e., we have $f'(N) = f(N, k_o)$. Hence, $f' \leq h$, and so certainly the inequality of the assertion will be satisfied. This candidate for a counterexample was not much more promising. Any time the structure of P is "check to see if k is the largest integer $\leq (N-1)$ such that . . . ", then P' will overhear this strategy, and proceed to find the needle directly by starting at $(N-1)$ and working down. Similarly for "smallest"; and any other "-est" that P' can figure out how to exploit.

The third example is the following. For N any positive integer, denote by p_N the integer obtained by writing out the digits of π (314159265...), and stopping as soon as you arrive at the largest integer less than N^2. [For example, $p_{32} = 314$.] Now let $\pi(N, k)$ be "yes" if and only if either: i) p_N is not the product of exactly two primes, and $k = 0$, or ii) p_N is the product of exactly two primes, and k is the smaller prime factor of p_N. In other words, the needle, for haystack N, is the smaller prime factor of p_N if p_N is a product of exactly two prime factors, and "0" otherwise. [The idea here is that the digits of π are "pretty random"; and that it is, presumably, hard to find prime factors other than by trial-and-error.] Now, most of the time (i.e., for most N) p_N will have many factors, and in these cases it will be easy to find the needle. Program P' will have a field day in these cases, easily achieving $f'(N) \leq \sqrt{N}h(N)$. But every so often (at least, we *hope* so — we, of course, have no theorem to this effect) p_N will turn out to be a product of two primes, and now the needle is harder to find. In this case, program P will have a relatively easy job of it: Given candidate k, P need only test to see whether or not k divides p_N. The shortcut program P', by contrast, has the duty to find the needle for this N — and it is hard to see how P' is going to do this other than testing various k to see if they divide p_N. So, here is an example in which (at least, *sometimes*) there doesn't appear a viable shortcut over the method of trial-and-error. So, is this a counterexample to the assertion? Probably not. The problem is that P must do more than merely check whether k divides p_N — it must also check whether or not p_N is a product of exactly two primes (in order to know whether or not $k_o = 0$). The difficulty (for P) of doing this is comparable to the difficulty

P' experiences in finding the needle in this case.

Either the assertion above is true; or it is false. It would be of great help in thinking about this subject, in my opinion, if we knew which. Indeed, as far as I am aware, we do not even have a counterexample to the stronger assertion that results from replacing the inequality with "$f'(N) \leq h(N)$".

17 Quantum-Assisted Computing

The discussion of the previous two sections suggests that the use of quantum mechanics may indeed gain efficiency for certain computations. But there remain at least three issues. First, as discussed in Sect. 16.4, our ability to apply quantum mechanics to specific problems appears to depend on finding a suitable technique for converting conventional computer programs to unitary operators. By "suitable", we mean a technique that results in no substantial loss of efficiency, and is such that no entanglements are created with any scratch Hilbert spaces that must be introduced. Second, we must make allowance for the fact that, while programs act on arbitrary strings (and thus are suitable for computing real problems), our unitary operators always act on *finite* tensor products of H's. And, finally, we must find a suitable definition of "difficulty" for unitary operators. We now introduce a general framework for computations using quantum mechanics, a scheme that, among other things, addresses these three issues.

Fix, once and for all, the following objects: i) a finite-dimensional Hilbert space H, ii) a unit vector $|\psi_o\rangle$ in H, iii) a finite list of unitary operators, each of which acts on some finite tensor product, $H \otimes \cdots \otimes H$, of H's, and iv) a finite list of projection operators[15], each of which acts on some finite tensor product of H's. The individual unitary and projection operators in these lists may operate on tensor products with different numbers of H-factors, e.g., some may act on a single H, some on $H \otimes H$, etc. We label each unitary operator of ii) and each projection operator of iii) by a nonempty string (e.g., as U_S and $P_{S'}$, respectively); and, for later convenience, we do not use the same string to label both a unitary and a projection operator. [We shall later impose a further condition on this arrangement; but for the moment it is convenient to keep things general.]

[15] A *projection* operator, P, is self-adjoint operator satisfying $P \circ P = P$, i.e., a self-adjoint operator having no eigenvalues other than 0 and 1.

We now introduce some terminology. First, we introduce a separator-character, $*$, in the manner we have done, occasionally, before. Next, we call a string $\tilde{S}$ a *unitary operation* if it is of the form $\check{S} * S_1 * \cdots * S_k$, i.e., consists of $(k+1)$ strings (each nonempty and containing no $*$; and with $S_1 \cdots S_k$ distinct), such that: The first of these strings, $\check{S}$, labels some unitary operator, $U_{\check{S}}$, in our list, and that $U_{\check{S}}$ acts on a tensor product of precisely k factors of H. Thus, beyond the first string, it is only the *number* of additional strings, and not what those strings are, that counts. For example, if, among the unitary operators, there is one labeled U_{8k}, and if it acts on $H \otimes H$, then "$8k * yzr * 8\$9Q$" would be a unitary operation; whereas "$8k * yzr$" would not. And, similarly, we call string $\tilde{S}$ a *projection operation* if it is again of the above form, $\tilde{S} = \check{S} * S_1 * \cdots * S_k$, such that $P_{\check{S}}$ appears on our list of projection operators, and it acts on the tensor product of exactly k factors of H. Note that no string is both a unitary operation and a projection operation; and that the problem of deciding whether a string $\tilde{S}$ is a unitary operation, a projection operation, or neither, is computable and has difficulty function $L(\tilde{S})$.

We now introduce a new computer language. As before, we have storage locations, each of which is labeled by a string and each of which contains a string (where, as before, $C(S)$ denotes the string contained in location S). There is a total of seven commands in this language, consisting of the five we introduced in Sect. 12 — INPUT, OUTPUT, APPEND, DELETE, and IF — together with two new ones:

6. APPLY $C(S)$.
7. OBSERVE $C(S)$, APPEND RESULT TO $C(S')$.

where, as before, S and S' denote arbitrary strings. Here is what these commands "do". In addition to the storage locations, there will now be a separate quantum system. The Hilbert space, $\mathcal{H}$, of states of this system will be, at any one moment during the operation of the computer, some tensor product of H's, where each factor of H in this tensor product is labeled by a string. Thus, we might have, at one moment, $\mathcal{H} = H_8 \otimes H_{abc} \otimes H_{Q3}$, a tensor product of three copies of H. The state of this quantum system, at that moment, will be given by some vector, say $|\Psi\rangle$, in the Hilbert space $\mathcal{H}$. Now, here is what is to be done in response to the command APPLY $C(S)$:

1. If $C(S)$ is not a unitary operation, then do nothing.

2. If $C(S)$ ($= \check{S} * S_1 * \cdots * S_k$, say) *is* a unitary operation, and each of the strings $S_1, \cdots, S_k$ is already represented by an H-factor in the tensor product that is $\mathcal{H}$, then apply to the state $|\Psi\rangle \in \mathcal{H}$ the unitary operator $U_{\check{S}}$ on $H_{S_1} \otimes \cdots \otimes H_{S_k}$. [That is, the unitary operator that is applied to $\mathcal{H}$ is the operator $U_{\check{S}}$ applied to the factors $H_{S_1} \otimes \cdots \otimes H_{S_k}$, and "$I$" applied to the remaining factors.]

3. If $C(S)$ ($= \check{S} * S_1 * \cdots * S_k$, say) is a unitary operation, and some of the strings $S_1, \cdots, S_k$ are not represented by H-factors in the tensor product that is $\mathcal{H}$, then proceed as follows. First, enlarge $\mathcal{H}$ to include those H-factors (i.e., replace $\mathcal{H}$ by its tensor product with the missing H_S). Next, replace the state $|\Psi\rangle$ by the result of taking the tensor product of this state with one copy of $|\psi_o\rangle$ for each new H-factor introduced. And finally, apply $U_{\check{S}}$ to this state in $\mathcal{H}$ (so enlarged) as in instruction 2.

Here is an example of these rules. Let, at some moment, $\mathcal{H} = H_{19} \otimes H_{yzr}$, let the state be $|\Psi\rangle \in \mathcal{H}$, let our list of unitary operators include an operator U_{8k} that acts on $H \otimes H$, and let $C(S) = 8k * yzr * Q9$. Then APPLY $C(S)$ would replace this $\mathcal{H}$ by $H_{19} \otimes H_{yzr} \otimes H_{Q9}$ and state $|\Psi\rangle$ by $|\Psi\rangle|\psi_o\rangle$; and would then apply $I \otimes U_{8k}$ to this state.

Similar rules apply to OBSERVE $C(S)$, APPEND RESULT TO $C(S')$. If $C(S)$ is not a projection operation, do nothing. Otherwise, proceed as follows. First, enlarge $\mathcal{H}$ by including, as necessary, additional H-factors, labeled by those strings in $C(S)$ not already so represented. Next, replace the state $|\Psi\rangle$ by that state in this enlarged Hilbert space obtained by taking one tensor product with a $|\psi_o\rangle$ for each new factor of H. Then, make an observation on this new state of this expanded Hilbert space, via the self-adjoint operator $P_{\check{S}}$. The result of this observation must be either 0 or 1, since $P_{\check{S}}$ is a projection. Append this result (suitably encoded, if necessary) to the string in location $C(S')$. [Usually, we would have previously SET $C(S') = \emptyset$, to avoid clutter.] After this OBSERVation, the state of our quantum system will, of course, be replaced by its projection into the appropriate eigenspace, i.e., by either $P_{\check{S}}$ or $(I - P_{\check{S}})$ applied to that state, according to whether the observation resulted in 1 or 0, respectively. It is generally more convenient to enlarge $\mathcal{H}$ using the APPLY command, rather than the OBSERVE.

In physical terms, what we are doing here is quite simple. We have some basic quantum system, described by Hilbert space H. What we call

a projection operation, for example, is a string that describes which projection operator is to be applied, and to which combination of copies the basic system. If any of the required copies are missing, we simply supply them, by ordering new copies of that system (which come with state $|\psi_o\rangle$ preinstalled) through the catalog, and placing those new system-copies next to the old system-copies. When all the necessary copies of our basic system have been assembled, we observe via the appropriate (projection) operator, and append the result to location S'. We remark that no generality has been lost by our making all observations via projection operators (rather than the more general self-adjoint operators). This is a consequence of the following fact: Every self-adjoint operator A on a finite-dimensional Hilbert space can be written as a sum, $a_1P_1 + \cdots + a_sP_s$, where the a_i are the eigenvalues of A, and the P_i project into the corresponding eigenspaces (so, in particular, the various P_i commute with each other). By virtue of this fact, we may, instead of observing via self-adjoint A, observe via each the P_i, noting that, by commutativity, the order of the latter observations is irrelevant. These two operations will always produce the same result, in terms of the outcomes and their probabilities as well as the final state of the system.

A *quantum-assisted program* is a finite, ordered list of commands, the first command of which is INPUT and the last of which is OUTPUT, such that each of these two commands appear nowhere else in the program. We run a program just as before, starting with each storage location containing the empty string; and with $\mathcal{H}$ the (one-dimensional Hilbert space of) complexes. We then execute the commands in order, except as directed by IF. The various commands will then manipulate strings; or operate on, expand, or observe $\mathcal{H}$. So, for example, the first APPLY or OBSERVE command will require that $\mathcal{H}$ be expanded to include the appropriate copies of H as a tensor product. If and when the program reaches OUTPUT, it halts, allowing us to read the output string.

When a given quantum-assisted program is run with a given input string, it must either halt, with some output string returned, or never halt (which we denote, as before, by "$*$"). From the laws of quantum mechanics, there will be probabilities for these various outcomes, i.e., we will have a probability distribution, p, on $\mathcal{S} \cup \{*\}$. We say that a quantum-assisted program *computes* problem π if, for every input string S, the program, run on that string, has $p(*) = 0$, and $p(\pi(S)) > p(S')$ for every $S' \neq S$. That is, the probability of not halting must be zero (although, as we have already seen,

it may still be possible that the program fail to halt), and the probability of the "correct answer", $\pi(S)$, must exceed that of every other possible output.

As an example of all this, let us consider the Grover construction. Let H be the two-dimensional Hilbert space of spin states of an electron, and let $|\psi_o\rangle$ be the state we earlier designated $|1\rangle$. Let U_1 act on H by $U_1|1\rangle = \frac{1}{\sqrt{2}}\{|1\rangle + |0\rangle\}$, $U_1|0\rangle = \frac{1}{\sqrt{2}}\{|1\rangle - |0\rangle\}$; U_2 by $U_2|1\rangle = -|1\rangle$, $U_2|0\rangle = |0\rangle$; and U_3 by $U_3|1\rangle = i|1\rangle$, $U_3|0\rangle = |0\rangle$. [A few more U's on H might also be required.] Let U_4 act on $H \otimes H$ by $U_4 = \exp(i(\pi/2)\vec{\sigma_1} \cdot \vec{\sigma_2})$. Finally, let there be a single projection operator, P_5, acting on a single H via $P_5|1\rangle = |1\rangle$, $P_5|0\rangle = 0$.

Our program will accept as input a positive integer n. Let us now introduce three subroutines. The first places, in some location S, the string "$1 * 1$", then executes APPLY $C(S)$, then places "$1 * 2$" in location S, then executes APPLY $C(S)$ again, and so on up to a total of n times. The result of running this subroutine is to set up our Hilbert space, $H_{\rm in} = H_1 \otimes \cdots \otimes H_n$, with each H_i-state given by $\frac{1}{\sqrt{2}}\{|1\rangle + |0\rangle\}$. [Note how we set up initial states, by starting with $|\psi_o\rangle = |1\rangle$ and applying the appropriate operator.] This is the set-up subroutine: It arranges the initial configuration. The next subroutine applies to this $H_{\rm in}$ the operator $V_{\rm in}$ followed by the operator $-W$. We are assuming that we have already been given the subroutine for $V_{\rm in}$, i.e., we have been given some rendering of our original needle-check program as a quantum-assisted program. The application of $V_{\rm in}$ may require an auxiliary Hilbert space, for scratch work). But we are demanding that in the final state the $H_{\rm in}$ and the scratch-work states not be entangled. Our earlier discussion of the construction of the operator W translates immediately into a quantum-assisted program, in our language, that applies this operator to $H_{\rm in}$. Finally the third subroutine observes $H_{\rm in}$ via the k-basis. This will, e.g., set $C(S) = 5 * 1$, execute OBSERVE $C(S)$, APPEND RESULT TO $C(S')$ and multiply the result (in $C(S')$) by 2^{n-1}; then set $C(S) = 5 * 2$, execute OBSERVE $C(S)$, APPEND RESULT TO $C(S')$, multiply the result (in $C(S')$) by 2^{n-2} and add to the previous result; and so on, for n times around. This subroutine, then, will end up producing some k-value, where $0 \leq k \leq N-1$, stored in some memory location.

Now, there are at least three different quantum-assisted programs that could be constructed from these subroutines. For the first, run the set-up subroutine, then the subroutine $-WV_{\rm in}$ once, then the observe subroutine, and finally report the k-value that results. For this program, we have $p(*) = 0$

(and, in fact, outcome $*$ would be impossible), $p(k_o) = 9/N$ (for large N), with the probabilities for the other k-values correspondingly reduced. This quantum-assisted program indeed computes our problem. For the second program, run the set-up subroutine, then the subroutine $-WV_{\text{in}}$ the correct number (approximately $(\pi/4)\sqrt{N}$) of times, then the observe subroutine, and finally report the k-value that results. For this program, we would again have $p(*) = 0$ (and, again, outcome $*$ would be impossible). Now, however, $p(k_o)$ is nearly one (for $p(k_o) > 1-1/N$), while the remaining k's have probabilities nearly zero. This quantum-assisted program also computes our problem. For our final program, we first do the computation of the previous program, but instead of reporting the k that results from the observation subroutine, we instead run the (non-quantum) needle-check program on that k, to find out if is indeed the needle. If it is, we report that k-value. Otherwise, go through the entire procedure again, finding a new k and checking that one. Continue in this way until we find the k that checks out as the needle. Note that, for this program, we must assemble a fresh copy of H_{in} for each running of the prior program, for that program results in a final H_{in}-state different from the initial state needed by that program. For this program, we again have $p(*) = 0$, but now the outcome $*$ is possible (for we could to go on indefinitely, being unlucky time after time). We also have $p(k_o) = 1$; and $p(k) = 0$ for all other k. Thus, this quantum-assisted program also computes our problem. Thus, we have three separate quantum-assisted programs, each of which computes the problem (whatever it is).

In these examples, the OBSERVE commands generally come after the APPLY commands have already been executed. But that, of course, needn't be the case in general: These commands could very well be intermingled. Note that the brain of this program is the original five commands and the storage locations with which they interact: These keep track of what is going on, decide when APPLY and OBSERVE are to be carried out, decide what to do with the results, handle the INPUT and OUTPUT, etc. The Hilbert space of the quantum system serves as a glorified storage register, into which we may place data (via APPLY), within which we may manipulate data (via APPLY), and from which we may extract data (via OBSERVE). We could also arrange, in effect, for there to be several different Hilbert spaces to handle data. These would be represented by different parts of the tensor product that is $\mathcal{H}$, and we would simply manipulate and access whatever part we wish to use at any one time, ignoring (i.e., applying the identity to, and not observing)

the other parts. At any one moment in the program, the Hilbert space of states of the quantum system, $\mathcal{H}$ is a finite tensor product of H's; although, of course, the number of H's in this product is not limited. Had we, for example, replaced each "$C(S)$" in the APPLY and OBSERVE commands with "S", then there would, for any given program, be a limit (independent of the input string) on this number. Replacing each "$C(S)$" in these commands by "$C(C(S))$" would not make any difference, since we have a subroutine SET $C(S') = C(C(S))$. Note that the structure of quantum-assisted programs does not require, necessarily, that entanglements be avoided; or that any entanglements that have been created eventually be undone. We simply APPLY certain unitary operators and OBSERVE certain projection operators; and whatever follows follows, whether there is entanglement or not. [Of course, it may be necessary to avoid entanglement in order that the program compute what we want it to compute.]

One point about this language should be emphasized. The quantum elements — the Hilbert space H, the initial state $|\psi_o\rangle$, the unitary operators, and the projection operators — are to be specified in full right at the beginning (once we know what problem it is we are to compute). It is not permitted to change these objects, depending on the input string.

18 Quantum-Assisted Computability

We introduced, in the previous section, the notion of a problem being computable by a quantum-assisted program; and we already have, from Sect. 6, the notion of a problem being computable by a regular program. Clearly, every regular-computable problem is quantum-assisted-computable (since every regular program *is* already a quantum-assisted program, just one happening to have no APPLY and no OBSERVE commands). Is the converse true? That is, is it true that every problem that is computable with quantum-assist is also computable in the regular sense?

This converse, as we have stated it above, turns out to be false. Here is an example. Let the Hilbert space H be 2-dimensional, with basis $|\psi_o\rangle$ (the initial state) and $|\psi_o\rangle^\perp$. Let there be just one unitary operator in the list, acting on a single copy of H by $U_\theta|\psi\rangle = |\psi_o\rangle \cos\theta + |\psi_o\rangle^\perp \sin\theta$, $U_\theta|\psi_o\rangle^\perp = |\psi_o\rangle^\perp \cos\theta - |\psi_o\rangle \sin\theta$, where θ is some fixed number. Let there be just one projection operator in the list, also acting on a single copy of H, by

$P|\psi_o\rangle = |\psi_o\rangle, P|\psi_o\rangle^\perp = 0$. [You will note that this is a pretty poor excuse for quantum-assistance: There are no operators in our lists that act on two or more H's, and so we will never produce, by means of these operators, entangled tensor-product states.] The program accepts as input a positive integer n, and proceeds as follows. It first executes APPLY U_θ to H_1; OBSERVE P on H_1, and accumulate the result (0 or 1) in $C(a)$. It then repeats this subroutine, with "H_1" replaced by "H_2", and so on, all the way up to "H_n". At this point, $C(a)$ will contain an integer (the number of times "1" was returned by OBSERVE during all n runs of the subroutine). Finally, the program returns, via OUTPUT, the rational number $C(a)/n$. What is this program doing? Well, on each run through the subroutine, either 1 will be added to the integer in $C(a)$ (probability $\cos^2\theta$), or it will not (probability $\sin^2\theta$). Thus, what the program returns at the end is a Monte-Carlo estimate of the value of $\cos^2\theta$, based on these n runs. The fractional error in this estimate goes down, as the number n of runs increases, as $1/\sqrt{n}$.

What this program does, in other words, is compute the number $\cos^2\theta$, in the sense of Sect. 7. What we have shown, then, is that for any number θ we can write a quantum-assisted program that computes $\cos^2\theta$. Now choose for θ that value such that $\cos^2\theta = c$, where c denotes the noncomputable number given by Eqn. (**??**) of Sect. 7. Thus, we have produced a quantum-assisted program that computes a number that is not computable with any regular program.

The discussion of the previous two paragraphs will fool nobody. It is absurd to take seriously a unitary operator U_θ that claims to carry out a rotation through a noncomputable angle: We would not expect to be able to buy and operate a machine that would apply any such U_θ to any real quantum system. Suppose, for example, that we acquire a machine that is capable of carrying out a rotation in the $|\psi_o\rangle - |\psi_o\rangle^\perp$ plane through any angle θ, where the value of θ is set by adjusting a knob. Then, as we fine-tune this machine, we shall (in order to know where to set the knob) be called upon to determine, more and more precisely, what the number θ is. That is, we shall have have to decide whether or not more and more complicated Turing machines will halt. One of these, for example, is the machine that searches for a counterexample to the Goldbach conjecture — and so, at this point, the proper adjustment of the knob will require that we settle this conjecture, one way or the other. And, as soon as we are finished with this one, we will be asked to resolve some other, even harder, conjecture in mathematics. How, in

this atmosphere, are ever going to get this experiment finished? It is silly to call this quandary a "piece of laboratory equipment". Note that the fact that there are rational numbers arbitrarily close to the noncomputable number c is of no help to us here. The issue is one of *determining* what c is (in practice, in the lab), not one of *approximating* c (in principle, in mathematics-land). To know c within 1%, let us say, will tell us whether or not the Goldbach conjecture is true. How helpful, in this circumstance, is it to be reassured that there does indeed exist a rational number within one-tenth of 1% of c?

> **Exercise.** Consider the unitary operator U_θ above, but let us select the angle θ "randomly, by spinning, and then stopping the knob". Then the probability is one that we shall end up with a noncomputable number (since the computable numbers in $[0, 2\pi]$ are measurable, with measure zero). So here is an example in which a quantum-assisted computer computes a (regularly-) noncomputable number. Respond.

So, we might ask, which unitary operators are, and which are not, "physically realistic"? In fact, the problem goes a little deeper than even this question suggests. Consider, for example, two unitary operators, U and U', each of which carries out a 90^o rotation in the (3-dimensional, say) Hilbert space H, but such that the planes in which these rotations take place make angle c with each other. Although each of these two unitary operators, by itself, is quite innocent-looking, together they allow, in the same manner as above, a computation of c. The lesson here is that we cannot look at the state $|\psi_o\rangle$, the unitary operators that appear on our list, and the projection operators that appear on our list, in isolation. It is the *entire list* — consisting of $|\psi_o\rangle$, the U's and the P's; all taken together — that must rise or fall. So, we might ask, how do we identify which such lists are, and which are not, physically realistic?

Here is a possible answer to this question. Let us imagine that the box in which the Hilbert space H is shipped has printed on it a standard basis for this Hilbert space, to be used for reference purposes. Then from this basis we acquire a standard basis for each tensor product, $H \otimes \cdots \otimes H$. Now, before purchasing a unitary operator U (on some tensor product of H's), we want to know what it is we are buying. This is to be provided by the manufacturer, in the form of a program, printed in the owner's manual, that accepts as input

any positive integer n, and returns rational numbers each within $1/n$ of the respective matrix element of U in this standard basis. We call this a program that *computes* U. Without such a program, we simply don't know what U "is". We have, similarly, the notion of a program that *computes* a projection operator (on a tensor product of H's), and of a program that *computes* the initial state $|\psi_o\rangle$. We now regard H, state $|\psi_o\rangle$, and the lists of unitary and projection operators as "physically realistic" if, for some H-basis, there exist programs that compute all these objects. This appears to be a rather mild condition.

Some terminology will allow us to formulate the idea of the previous paragraph more neatly. Fix, as before, i) a finite-dimensional Hilbert space H, ii) a unit vector $|\psi_o\rangle$ therein, iii) a finite list of unitary operators (labeled by strings) on various H-tensor products, and iv) a finite list of projection operators (labeled by strings) on various H-tensor products. We call a string **S** a *history* if it is of the form $S_1 * *S_2 * * \cdots * *S_k$, where each of the strings $S_1, \cdots, S_k$ is either i) a unitary operation (string), or ii) a projection operation (string) to which either "$*0$" or "$*1$" has been appended. Thus, a typical history string might be "$k9 * yzr * 0 * *B * xx * ABC$": The rightmost entry represents the unitary operator U_B applied to $H_{xx} \otimes H_{ABC}$; the other entry the projection operator P_{k9} applied to H_{yzr}; with "$*0$" appended.

Now consider the running of some quantum-assisted program. Each time an APPLY command is executed, some things will be done with respect to the Hilbert space $\mathcal{H}$: This Hilbert space will be expanded (if necessary) by taking additional tensor products with H-copies; the state $|\Psi\rangle$ will be adjusted (if necessary) with $\otimes|\psi_o\rangle$'s to lie in this expanded Hilbert space; and a certain unitary operator will be applied to this state. The same goes for an OBSERVE command, except that in this case either the projection operator P (from the list) is applied to the state (the case in which the observation yielded "1"), or the projection operator $(I-P)$ is applied (observation yielded "0"). The other commands, INPUT, OUTPUT, DELETE, APPEND, and IF, do nothing with respect to $\mathcal{H}$. Thus, as of any one moment during the running of the program, $\mathcal{H}$ will have been subjected to some finite number of such operations, in some order. But this is precisely the information contained in a history. In other words, a history string provides a complete summary of what has happened with respect to $\mathcal{H}$ as of a certain moment. Perhaps "virtual history" would be a better term, for we admit as histories *all* strings formed by the rules of the previous paragraph, whether or not they

happen to represent what has actually occurred. From the history string we may determine what the Hilbert space $\mathcal{H}$ is at that moment, what state (in $\mathcal{H}$) the quantum system is in, and what information (0's and 1's) has been passed so far from the quantum system to storage locations. From the history "$\mathbf{S} = k9 * yzr * 0 * *B * xx * ABC$" of the previous paragraph, for example, we would determine that $\mathcal{H} = H_{yzr} \otimes H_{xx} \otimes H_{ABC}$, that the state is $((I - P_{k9}) \otimes I \otimes I)((I \otimes U_B)(|\psi_o\rangle|\psi_o\rangle|\psi_o\rangle))$, and that value "0" was returned on the execution of the one OBSERVE command. The idea, in short, is to reflect the entirety of the quantum part of quantum-assisted computing by a string, something we can easily dissect.

For $\mathbf{S}$ any history, denote by $\gamma(\mathbf{S})$ the squared norm of the state (as determined by $\mathbf{S}$) in the Hilbert space (as determined by $\mathbf{S}$). Thus, for example, we have $\gamma(\mathbf{S}) = 1$ if the history $\mathbf{S}$ contains no projection operations (since $|\psi_o\rangle$ is unit, and unitary operators are norm-preserving); and, for a general history $\mathbf{S}$, $0 \leq \gamma(\mathbf{S}) \leq 1$. This real-valued function γ on histories carries all the relevant information about the workings of quantum mechanics within our quantum-assisted language, in a sense that will become clear shortly.

Now suppose that, with respect to some standard basis for H, there exists a program that computes the initial state $|\psi_o\rangle$, as well as ones that compute each of the unitary and projection operators in our lists, in the sense described above. Then we may combine these to produce a program that, given any history $\mathbf{S}$, will compute the components, in terms of this standard basis, of the state determined by this $\mathbf{S}$. It is now a simple matter to write a program that computes the function γ, in the following sense. This program accepts as input any history $\mathbf{S}$ and positive integer n, and returns some rational number within $1/n$ of $\gamma(\mathbf{S})$. The existence of such a program implies, of course, that each number $\gamma(\mathbf{S})$ is computable, but it also implies a great deal more: It means that there is a *single* rule that suffices to provide an approximation for *every* $\gamma(\mathbf{S})$.

We now return to the issue raised at the beginning of this section. We claim: Any problem that is computable by a quantum-assisted program, using computable initial state and operators, is also computable by a regular program. The proof, much like that of the similar result for probabilistic programs, is by simulation.

Fix H, and let $|\psi_o\rangle$ and the labeled unitary and projection operators be computable, in the sense above. Now fix a quantum-assisted program P_{quant},

together with the input string, S_{in}, on which this program is to be run. We are going to construct a regular program, P_{reg}, that will simulate the running of P_{quant} on S_{in}. Suppose that P_{quant} has been run for a few steps, encountering an APPLY command or two, but no OBSERVE command. Then the entire state of the computer (including the quantum system) can be expressed, as of this moment, by giving three pieces of information: i) which command in the list P_{quant} is slated to be executed next; ii) which string is stored in each (nonempty) storage location; and iii) the history, $\mathbf{S}$, of the quantum system. Let us now carry P_{quant} through the next step (say, a non-OBSERVE one). To simulate this step, we update the three pieces of information in the obvious way: For i), we now indicate what is the new next command; for ii) we make an adjustment (required only if that step was APPEND or DELETE; and then to only one of the stored strings) to reflect the new string-storage; and for iii) we add an entry (required only if that step was APPLY) to the history. In this way, we continue our simulation of P_{quant}, step by (non-OBSERVE) step. What happens when we reach an OBSERVE command? Now there will be two possible outcomes, depending on whether the observation returns 1 or 0. We reflect this state of affairs by splitting our description into two branches, each of which carries three pieces of information as above. In one branch, corresponding to the observation returning "1", the three pieces of information read: for i), the next command to be executed; for ii), the same stored strings, but with "1" appended to one particular string; and, for iii) the same history string, but with the addition a certain projection operation and " $*1$". In the other branch, we enter, similarly, the three pieces of information appropriate to the case in which observation returns "0". We now continue to simulate the behavior of P_{quant} in each of these two branches separately. As more OBSERVE commands are executed, the number of branches will grow, as will the burden of separately simulating what happens within each branch. But at every stage in this simulation, we shall have a finite number of branches, each described by just these three pieces of information.

So, P_{reg} simulates P_{quant}, in this way. Every so often, one of the branches being simulated by P_{reg} will reach an OUTPUT command (after which there is nothing more to simulate). When that occurs, the program P_{reg} reads the output string S and the final history $\mathbf{S}$ for that branch, stores this information in a special section, and drops that branch from further consideration. [We, of course, know that $\gamma(\mathbf{S})$ is the probability that the actual program P_{quant}

will take this branch, reaching this OUTPUT and returning this S.] Thus, as the simulation by $P_{\rm reg}$ continues, the list of string-history pairs in this special section will grow. Our program $P_{\rm reg}$ will include, furthermore, a subroutine, which accepts as input a positive integer n, and operates as follows. The subroutine goes to this special section, takes each of the history strings in that section, and computes γ of that history, within error $1/n$ (here using the program that we constructed from the reference manuals). It then totals these numbers, for each output string listed in that section. Finally, the subroutine checks to see if any one output string, say S, has emerged as a clear winner (i.e., is such that no other string S' will ever be able to achieve a total exceeding that for S, even if we allocate to S' all the so-far unallocated probability, and even if we assume that all the $1/n$-errors in these probabilities are resolved in S''s favor). If the subroutine finds such a clear winner S, then it causes $P_{\rm reg}$ to halt immediately, giving that S as output. If there is no clear winner, then the subroutine returns $P_{\rm reg}$ to its simulation.

The full program $P_{\rm reg}$ now operates as follows: Every so often (say, after every hundred steps of simulation), $P_{\rm reg}$ runs the subroutine, using an n-value one higher than that for the previous run. So, this program $P_{\rm reg}$ will continue to run in this way: continuing to carry out the simulation of $P_{\rm quant}$, continuing to store the results for halted branches in the special section, and continuing to make ever-finer checks on the status of that special section. Now suppose that the program $P_{\rm quant}$ computes a problem π, i.e., that, for every input string $S_{\rm in}$, $P_{\rm quant}$ has probability zero of failing to halt, and has probability of halting with output $\pi(S_{\rm in})$ that exceeds that of every other possible output. In this case, our simulation $P_{\rm reg}$ must eventually halt, for eventually it will have accounted for sufficient probability to identify $\pi(S_{\rm in})$ as the clear winner. At this point, $P_{\rm reg}$ will return $\pi(S_{\rm in})$.

What we have shown, then, is that, given any quantum-assisted program that computes a problem, we can, by simulating it in this manner, build a regular program that computes the same problem. Note that $P_{\rm reg}$ *always* halts, giving the correct answer, $\pi(S_{in})$. The quantum-assisted program $P_{\rm quant}$, by contrast, gives this answer only probabilistically. Note also that this argument makes essential use of the program that computes the function γ. In any case, we conclude that any problem that is computable by a quantum-assisted program, with computable $|\psi_o\rangle$ and operators, is also computable by a regular program.

The result of this section is of perhaps only mild interest. What is impor-

tant, and what we shall use extensively in the what follows, is three notions: that of a history; that of the function γ; and the present strategy for simulating a quantum assisted computation.

19 Quantum-Assisted Difficulty Functions

In Sect. 10, we defined a difficulty function for every program that computes some problem, π. This positive, real-valued function represents the amount of "computer time", as a function of the input string $S_{\rm in}$, required to compute $\pi(S_{\rm in})$. We now wish to do the same thing for any quantum-assisted program that computes a problem. We shall do so in two steps. First, we assign a difficulty to each individual command in our quantum-assisted language. This will entail a certain restriction on the unitary and projection operators that go into the language. Second, we adapt the notion of a difficulty function to take into account the fact that quantum-assisted computing provides answers only probabilistically.

We emphasize again that we have set things up so that the quantum elements — the Hilbert space H, the initial state $|\psi_o\rangle$, and the unitary and projection operators – are to be specified as part of the program, i.e., may depend on the problem to be computed, but not on the particular input string. Imagine that one had done otherwise — e.g., had allowed the use of different unitary operators for different input strings. Then how would we produce any reasonable notion of a difficulty function? It could be, for example, that certain input strings would require unitary operators that are very delicate — that can be created only with a great deal of time and effort. But this "time and effort" will have to go into the difficulty for those input strings. That is, in order to produce a difficulty function in such a language we would have to quantify the delicacy of unitary operators. Such a project — to put it mildly — does not look easy. We remark that many of the "quantum computations" that have been proposed suffer from precisely this defect.

There are seven types of commands that may appear in a quantum-assisted program. Five of these — INPUT, OUTPUT, APPEND, DELETE, and IF — are the original commands we introduced in Sect. 12; and it is natural to assign to these commands the difficulties we already chose earlier. But what of the two new commands — APPLY and OBSERVE? Note that we

have finite lists of the unitary and projection operators that are permitted to appear in these commands. So, in effect, each of APPLY and OBSERVE represents a finite number of physical operations. A natural choice would thus seem to be: Assign, to each APPLY and OBSERVE command, difficulty one. But, unfortunately, things are not that simple, as the following example illustrates.

Consider a problem, π, that accepts as input a positive integer n, returning either 0 or 2; and is such that every program that computes this π has difficulty function f that grows very quickly with n (say, faster than 2 to the power of 2 to the power of 2, etc, for n iterations). We have seen in Sect. 11 that there does indeed exist such a (computable) problem. Now set $c' = \sum_n \pi(n)/3^n$, a certain real number. Thus, this c' is constructed in the same manner as the noncomputable number c of Sect 6, but, in contrast to that c, is a computable number (since the problem π is). The point, however, is that c' is hard to compute: The number of computer-steps required to approximate c' within $1/n$ grows very quickly with n. Let us now consider a quantum-assisted language in which the Hilbert space H is 2-dimensional, and the list of unitary operators includes a U_θ that applies a rotation, to a single H, through angle θ, just as in the previous section. Now, however, we choose $\cos^2\theta = c'$. We next write a quantum-assisted program in this language similar to that of the previous section. That is, this program repeatedly applies U_θ and makes an observation, resulting in a Monte-Carlo estimate of $\cos^2\theta$. In order to compute $\pi(n)$, we must estimate the value of c' to within $1/(2*3^n)$. But the error in a Monte-Carlo estimate decreases as the reciprocal of the square root of the number of runs. Thus, we need about $(2*3^n)^2 = 4*9^n$ applications of U_θ to have a reasonable chance of recovering the value of $\pi(n)$. For given n, carry out 10^{n+1} applications (just to be on the safe side): Then we shall have a probability of correctly determining $\pi(n)$ that is high and increasing with n. Thus, we have written a quantum-assisted program that computes this problem π with, presumably, difficulty function 10^{n+1} — much less than the difficulty function of any regular program that computes π.

This was a foolish argument in the previous section, and it is no better this time around. What this argument does show, however, is that we must be prepared to exercise some care as to which unitary (and projection) operators will be allowed in quantum-assisted programs, and as to what their difficulties are to be.

It is tempting to take the position that, since this U_θ is apparently such a terribly difficult operator, we can make things right by merely assigning a large difficulty to to the corresponding APPLY command. But this isn't going to work: There is just one of these U_θ's, and just one corresponding APPLY command, and so there is just a one number for us to assign. Changing the difficulty of this single command from 1 to 1000, for example, will not change the difficulty function (up to equivalence), and so will not undermine the above argument. In fact, it appears that any attempt to preserve this particular U_θ in our list of unitary operators will return us to the issue of how we take into account errors. What is problematic about this operator U_θ appears only in attempts to "approximate" it. Imagine a new kind of quantum-assisted program that, when commanded to APPLY this U_θ, actually applies an operator that is only a rough approximation to U_θ, but which (by virtue of the roughness of the approximation) is also low in difficulty. If and when, as the running of the program proceeds, a more accurate application of U_θ becomes necessary, then our program would go back and redo the original APPLY command, but this time applying something that is closer to the actual U_θ (and carries a larger difficulty).

So, it appears that the only way we can avoid a very complicated programming environment, in which errors must constantly be taken into account, is to banish this U_θ from out list of unitary operators. But this is a slippery slope: Will there be allowed other unitary operators, based on numbers that are a little easier to compute, but still pretty hard? Where do we draw the line? Our approach will be to go ahead and slide down the slope, i.e., to rule out all but the "simplest" U's.

Fix an m-dimensional Hilbert space H, a unit vector $|\psi_o\rangle$ in H, and finite lists of unitary and of projection operators, each acting on some finite tensor product of H's. We say that this arrangement is *simple* if, for every history string $\mathbf{S}$, the number $\gamma(\mathbf{S})$ is rational; and furthermore there exists a (regular) program that computes this γ, with difficulty function f satisfying

$$f(\mathbf{S}) \leq N_{\rm op}(\mathbf{S})\ m^{N_{\rm H}(\mathbf{S})}. \tag{13}$$

Here, $N_{\rm op}(\mathbf{S})$ denotes the total number of (unitary or projection) operations represented by the history string $\mathbf{S}$, $N_{\rm H}$ denotes the total number of copies of the Hilbert space H that appear in the final tensor product (within which $\gamma(\mathbf{S})$ is computed), and m is the dimension of H. Note that simplicity implies immediately that γ is computable, in the sense of the previous section.

Here is why this definition is what it is. Fix some basis for H, say $|\alpha_1\rangle, |\alpha_2\rangle, \cdots, |\alpha_m\rangle$. Then, as we have seen, we may construct from this H-basis a basis for every tensor product, $H \otimes H \otimes \cdots \otimes H$, of H's. For n H's in the tensor product, this basis will contain m^n vectors, each of which is a product of some n vectors in our H-basis, e.g., $|\alpha_i\rangle|\alpha_j\rangle \cdots |\alpha_k\rangle$. Now suppose that, with respect to this basis, the components of $|\psi_o\rangle$ and of all the unitary and projection operators in our list are rational numbers. This is about as simple as $|\psi_o\rangle$ and the U's and P's could possibly be. Now, in this case each value of γ will certainly be rational. Furthermore, we can easily write a program that computes γ. This program would take the history $\mathbf{S}$ and express explicitly, in terms of our basis, the result of applying in succession each operation contained in $\mathbf{S}$. The program then takes the resulting final state, again expressed in terms of components in this basis, and computes its squared norm.

What is the difficulty function of this program? Consider one of the operations — say, application of some unitary U — in the history string $\mathbf{S}$, and suppose that, at the point at which this U is applied, the Hilbert space is a tensor product of n copies of H. Then, at that point, the dimension of this Hilbert space will be m^n, and so the current state will have m^n components, and so the record of this state in our program will require that m^n entries be stored. To apply the operator U to this state will entail replacing each of these entries by a linear combination of other entries. That is, in order to compute the effect of this U we shall have to carry out a number of arithmetic operations given by a small multiple of m^n. Note — and this is a key point — that there is no savings from the fact that U actually operates on a small number of H's in that tensor product. For example, say that H has basis $|0\rangle, |1\rangle$, and let U act on a single H, by $U|1\rangle = \frac{4}{5}|1\rangle + \frac{3}{5}|0\rangle$, $U|0\rangle = \frac{4}{5}|0\rangle - \frac{3}{5}|1\rangle$. Let the current tensor product consist of a large number n of H's, and suppose that this U is acting on the 73^{rd} one. Consider any two basis-elements,

$$|0\rangle|1\rangle|1\rangle|0\rangle \cdots |1\rangle \cdots |0\rangle|1\rangle,$$

$$|0\rangle|1\rangle|1\rangle|0\rangle \cdots |0\rangle \cdots |0\rangle|1\rangle.$$

differing only in their entry for the 73^{rd} copy of H. Now, the action of U will mix these two elements. Thus, to compute how this U acts, we will have to carry out a small arithmetic computation involving the component-values stored in these two locations. But the same is true for all m^n component-values stored. So, the order of m^n arithmetic operations must be carried out.

And, apparently, there is available no shortcut, by which multiple entries can be calculated or stored all in one shot. The next application of a U may involve the 194^{th} H in the tensor product, and to compute its action will again involve the entries in *all* the m^n locations, grouping those entries in a different way from that of the previous application of U.

So, under the assumption of rational components in a certain basis, the difficulty required to compute the effect of application of one U or P in our list is a small multiple of m^n. So, the total difficulty to compute the rational number $\gamma(\mathbf{S})$ is a sum of terms, one for each operation in the history $\mathbf{S}$ and each of the form m^n, where "n" is the then-number of H's in the Hilbert space. Eqn. (**??**) is a simpler, and somewhat weaker, expression of this bound. We conclude: In the case in which $|\psi_o\rangle$ and the U's and P's have rational coefficients in some H-basis, that arrangement is simple as defined above. In fact, a few other cases are also allowed by the definition, e.g., that in which U is of the form $U|1\rangle = \frac{1}{\sqrt{2}}\{|1\rangle + |0\rangle\}$, $U|0\rangle = \frac{1}{\sqrt{2}}\{|1\rangle - |0\rangle\}$. The definition of "simple" as given has the advantage that it allows these other cases, and also that it makes no reference to any basis.

So, in short, a system — of $|\psi_o\rangle$, some unitary U's and some projection P's — is simple if the only thing that counts in computing $\gamma(\mathbf{S})$ is the number of operations represented by $\mathbf{S}$ and the size of the Hilbert spaces on which these operations act. There is no factor to represent "how hard" the arithmetic manipulations are. Simplicity means, in effect, that the operators require only "easy" arithmetic.

We are now in a position to appreciate the key difference between a quantum-assisted program and a regular program. The quantum-assisted program can apply one of its unitary or projection operators in a single step. This is because the operators themselves are rather simple, and each of them applies to only a few H's. The quantum-assisted computer simply assembles the appropriate two or three H's, and applies the operator — all without even knowing about any other H's that may be involved in the tensor product. But, in order for a regular program to see what is happening, it is necessary for that program to consider *all* the H's in the tensor product: It cannot simply ignore those H's to which the operator does not apply. In short, quantum mechanics is able to *do* (easily; probabilistically) what non-quantum mechanics can only *compute* (with much more difficulty; numerically). This state of affairs is reflected by the fact that the regular program ends up with a difficulty function for γ satisfying Eqn. (**??**), whereas the analogous

inequality for a quantum-assisted program would read: $f_{\rm quant}(\mathbf{S}) \leq N_{\rm ops}(\mathbf{S})$. What quantum mechanics has going for it, in short, is the tensor product.

So, we have decided what combinations of H, $|\psi_o\rangle$, U's, and P's (namely the simple ones) to allow in our quantum-assisted programs; and what difficulty (namely, one each) to assign to the new commands, APPLY and OBSERVE, in that language. We must now contend with the probabilistic aspect of quantum-assisted computing.

Fix a quantum-assisted program, $P_{\rm quant}$, that computes some problem, π, in the sense of Sect. 17. Thus, for every input string, $S_{\rm in}$, we have a probability distribution p on the possible outcomes with this $S_{\rm in}$; and these satisfy $p(*) = 0$, and $p(\pi(S_{\rm in})) > p(S')$ for every $S' \neq \pi(S_{\rm in})$. We wish to assign a difficulty function to this entire program. The situation here is similar to that we faced in our discussion, in Sect 13, of probabilistic computing. We must take account of the fact that different runs of our program may require different numbers of steps; and also that the output from running our program, on input string S, may be something other than the right answer, $\pi(S)$. And we adopt the same formula as we obtained earlier: We assign, for the difficulty for the computation on input string S, the value $f(S) = D(S)(p+p')/(p-p')^2$, where $D(S)$ is the expected difficulty in running this program on this input string; p denotes the probability of the correct output, $\pi(S)$, and p' denotes the probability of the next most probable output. The first factor on the right corrects for the fact that different runs of our program may encounter different difficulties — we take the mean difficulty. The second factor corrects for the probabilistic nature of the output.

As an example of these ideas, consider again the Grover construction, as reflected by the three distinct quantum-assisted programs introduced in Sect. 17. We now determine the difficulty function for each of these programs. For input n, a positive integer, denote by $h(n)$ the difficulty encountered by a quantum-assisted subroutine in applying the entire operator WV to the tensor product, $H \otimes \cdots \otimes H$, of n H's, so $h(n) \geq n$. Then, as part of the lore of this construction, this same $h(n)$ will be the largest difficulty encountered by a regular program making a check to see whether a single k is the needle in the n-haystack. Thus, a regular program can compute this problem with difficulty function $f_{\rm reg}(n) = Nh(n)$, where $N = 2^n$ is the total number of needles in the haystack.

In the first quantum-assisted program, there is made a single iteration of

WV, followed by a series of n OBSERVations. There results a candidate k for the needle, which is then immediately reported using OUTPUT. The probability (p) that this k is the actual needle is about $9/N$; while the remaining k-values share the rest of the probabilities (so each p' is about $1/N$). The mean difficulty in this case is $D(n) = h(n) + n$ (since there is performed a single iteration followed by n observations). Substituting into our formula, we obtain a difficulty function (for large N) $f_{\rm quant}(n) = (10/64)(h(n)+n)N$, which is equivalent to the difficulty function, above, of the regular program. It should come as no surprise that *this* strategy for a quantum-assisted computation brings no advantage.

In the second program, there is made a total of $\sqrt{N}$ (give or take a couple) iterations of WV, again followed by a series of n OBSERVations, and the reporting of a needle-candidate. Here, the mean difficulty is $D(n) = \sqrt{N}h(n) + n$. The probability that the reported k is the actual needle is now about $p = 1 - \frac{1}{N}$, while the remaining k-values share the rest of the probabilities (so p' is approximately $\frac{1}{N^2}$). Substituting into our formula, we obtain difficulty function $f_{\rm quant}(n) = (\sqrt{N}h(n) + n)(1 - \frac{1}{N} + \frac{1}{N^2})/(1 - \frac{1}{N} - \frac{1}{N^2})^2$. This function, for large N, is equivalent to $\sqrt{N}h(n)$. Note that the difficulty function for this program is $\ll$ than the difficulty function for the regular program, reflecting a potential advantage for the quantum-assist (which would, perhaps, be a real advantage, if only we had a good candidate for what problem is being computed here).

In the third program, we begin, just as above, with a total of $\sqrt{N}$ iterations of WV, followed by a series of n OBSERVation. But in this case we check, using the regular program, whether or not the k that results is indeed the needle. If it is, report that k (thus incurring total difficulty $\sqrt{N}h(n) + n + h(n)$); but if it is not, go back to the beginning, carrying out the iterations and the OBSERVations again. Repeat until you find the needle. In this case $p = 1, p' = 0$ (since we will either find the needle, or (with probability zero) continue trying forever). But now the mean difficulty (which really is a mean in this case, for now there is a nontrivial probability distribution on difficulties) is more complicated. The probability that we carry out just one group of $\sqrt{N}$ iterations of WV is $1 - \frac{1}{N}$ (approximately); that we carry out two is $\frac{1}{N}(1 - \frac{1}{N})$; etc. So, the mean difficulty is given by

$$D(n) = (\sqrt{N}h(n)+n+h(n))[1(1-\frac{1}{N})+2\frac{1}{N}(1-\frac{1}{N})+3\frac{1}{N^2}(1-\frac{1}{N})+\cdots]. \quad (14)$$

The sum on the right is $(\frac{N}{N-1})$. Substituting these into our formula, we obtain, for large N and up to equivalence, the difficulty function of this program: $f_{\text{quant}}(n) = \sqrt{N}h(n)$. This is identical to the difficulty function of the previous program.

These are precisely the results that we expect. The first program is not really exploiting the potential advantages of quantum mechanics, and its difficulty function shows it. The last two are essentially the same program. The only difference is that the first program has a fixed difficulty per run (as opposed to a probability distribution in difficulties), but leaves some unfinished business in form of the output-probabilities; while the second yields certainty for the correct output, at the cost of possibly requiring several repetitions. Our definition of the difficulty function of a quantum-assisted program is so constructed to ignore such window-dressing.

We remark that one way to exploit quantum mechanics in the computation process is to use it merely as a random-number generator. That is, there would not, in the course of the computation, be created any entanglements between the various H-factors in the tensor product. Instead, we would simply create an H-factor (using APPLY), with initial state $|\psi_o\rangle$), and then immediately thereafter make an observation on that factor (using OBSERVE), resulting in various probabilities for the various outcomes. These two steps would be repeated, as necessary, throughout the program. Indeed, every probabilistic program (as defined in Sect. 13) can in this manner be simulated by a quantum-assisted program. For the quantum-assisted programs in this class, each $\gamma(\mathbf{S})$ that will have to be evaluated can be computed with (up to equivalence) difficulty one (as follows from the fact that no entanglements are created between the H-factors). Note further that the difficulty function for a quantum-assisted program is virtually a rewrite of the difficulty function for a probabilistic program. There follows from these remarks:

Theorem. Let π be a problem, and $\mathcal{P}_{\text{prob}}$ any probabilistic program that computes that problem. Then there exists a quantum-assisted program, $\mathcal{P}_{\text{quant}}$, that also computes π, such that $\mathcal{P}_{\text{quant}}$ has precisely the same difficulty function as $\mathcal{P}_{\text{prob}}$.

In other words, any benefits that probability might bring to the computation process are also borne by quantum mechanics. In Sect. 13, we remarked that it is, apparently, an open question whether there exists a problem such

that some probabilistic program computes that problem more efficiently than any regular program. If there were such a problem, then it follows from the theorem that some quantum-assisted program would also compute that problem more efficiently than any regular program. Of course, the status of the converse of this assertion is not clear: It could turn out that there exists no problem for which probability increases efficiency, and yet there does exist a problem for which quantum mechanics increases efficiency.

20 Quantum-Assisted Efficiency I

This completes our formulation of quantum-assisted computing. This formulation begins by fixing a character set; together with a finite-dimensional Hilbert space H, a state in that Hilbert space, and finite lists of unitary and projection operators, each acting on some finite tensor product of H's. On these objects we impose the condition of simplicity. We then introduce a quantum-assisted programming language, consisting of some seven commands. We introduce the notion of a program's computing a problem; as well as the difficulty function associated with such a program. These are the building blocks of quantum-assisted computing. In this section and the next, we compare quantum-assisted programs and regular programs with respect to their difficulty functions. Here, we obtain a result to the effect that the maximum reduction in difficulty that can be achieved by quantum-assist is logarithmic.

Fix a quantum-assisted program, P_{quant}, that computes some problem π, and denote its difficulty function by f_{quant}. We construct a regular program, P_{reg}, that simulates P_{quant}, in the following manner. Fix the input string, S_{in}. Then P_{reg} simulates the running of P_{quant}, on this input string, in the same manner as in Sect. 18. That is, at any one moment P_{reg} is following a number of "branches" of P_{quant} (each spawned by the simulated execution of an OBSERVE command); and for each of these branches P_{reg} keeps track of three pieces of information: i) what is the next command, in the list P_{quant}, to be executed; ii) what is stored, by P_{quant}, in all nonempty storage locations; and iii) what is the history string $\mathbf{S}$, representing interactions P_{quant} has initiated with the quantum system. We now modify that earlier simulation, in two ways.

First, the earlier simulation (implicitly) proceeded along each branch at

the same command-rate. That is, one additional command was executed in every branch; then one more command in every branch, etc. Now, however, we proceed along each branch at the same difficulty-rate. That is, we carry out one unit of P_{quant}-difficulty in each branch; then one more unit in each branch, etc. Thus, branches that involve a great deal of difficulty per command are simulated more slowly than those that involve less.

For the second modification, recall that in the earlier simulation P_{reg} maintained in its memory a special section, which was added to each time a branch under simulation reached a "halt", i.e., a P_{quant}-OUTPUT command. When this occurred, the program P_{reg} stored in this section the current history string $\mathbf{S}$, as of that halt, as well as the string S that would have then been returned by P_{quant}. A branch, once reported in this way, was then abandoned by P_{reg}. The present simulation is a little different. The special section now contains a certain list of strings and, for each such string S, a corresponding rational number. When a branch, while under simulation, reaches a halt, P_{reg} immediately computes the rational number $\gamma(\mathbf{S})$ (where $\mathbf{S}$ is the current history string), adds this number to the number already stored for string S (where S is the string that P_{quant} would have returned), and then again abandons that branch. Thus, the various strings listed in this special section are, as before, the possible outputs from P_{quant} up to this point. But now the (rational) number stored for each string gives the total probability that P_{quant} would, by this point, have returned that string. In addition, P_{reg} contains a subroutine, which operates as follows: It goes through the list of strings and (rational) probabilities in the special section, and determines whether any output string in that list can be declared a clear winner (i.e., has a total that is greater than that which could be achieved by any other string, even if that string were allocated all so-far unallocated probability). If the subroutine finds a clear winner, then P_{reg} itself halts, returning the winning string. This subroutine is run each time P_{reg} finds itself making an addition to the special section.

So, given the program P_{quant}, we may write this program P_{reg}, which, for every input string, simulates the behavior of P_{quant}, as described above. Clearly, this P_{reg} always halts, and computes the same problem as P_{quant} does. Denote by f_{reg} the difficulty function of P_{reg}. The plan is to use (**??**) to find an inequality that bounds f_{reg} in terms of f_{quant}.

Fix the input string S_{in}, and denote by p the probability that P_{quant} will return $\pi(S_{\text{in}})$, and by p' the probability of the next-most-likely output, so

$p > p'$. Then $P_{\rm reg}$ will be able to declare a clear winner, and so will halt, at least by the time it has accounted for an amount $1 - \frac{p-p'}{2}$ of probability[16]. Denote by $\mathcal{N}$ the total amount of $P_{\rm quant}$-difficulty that $P_{\rm reg}$ has simulated (in each branch) at the point at which $P_{\rm reg}$ halts. Then we have

$$f_{\rm quant}(S_{\rm in}) = D(S_{\rm in}) \frac{p+p'}{(p-p')^2} \geq \{\mathcal{N} \frac{p-p'}{2}\} \frac{p+p'}{(p-p')^2} \geq \frac{1}{2}\mathcal{N}. \qquad (15)$$

The first step in (**??**) is the definition of $f_{\rm quant}$. The second step uses the fact that $P_{\rm reg}$ has already gone through amount $\mathcal{N}$ of $P_{\rm quant}$-difficulty, and yet there still remains probability at least $\frac{p-p'}{2}$ that $P_{\rm quant}$ has not halted. This fact alone contributes to the mean total difficulty of $P_{\rm quant}$ an amount equal to the product of these two numbers.

We must now relate $f_{\rm reg}$ to this $\mathcal{N}$. To this end, denote by $M_{\rm H}$ the maximum number of additional H's that can be introduced into the tensor product per unit difficulty. For example, if no unitary or projection operator in our original lists requires a tensor product of more than three H's, then we would have $M_{\rm H} = 3$. Note that $M_{\rm H}$ depends *only* on the our quantum-assisted language and *not* on the particular program under consideration. Returning now to our simulation, since $P_{\rm reg}$ has traversed total $P_{\rm quant}$-difficulty $\mathcal{N}$ in each branch, the total number of operators that have been applied in each branch is bounded by $\mathcal{N}$; while the total number of H's that can occur in the tensor product in each branch is bounded by $M_{\rm H}\mathcal{N}$. We have

$$f_{\rm reg} \leq \{2^{\mathcal{N}}\} \{\mathcal{N} + \mathcal{N}^2 \, m^{M_{\rm H}\mathcal{N}}\}. \qquad (16)$$

The first factor on the right is an upper bound on the total number of branches, where we are using the fact that each OBSERVE command spawns the splitting of one branch into two. The second factor on the right in (**??**) is an upper bound on the total $P_{\rm reg}$-difficulty of each branch. The first term therein covers the case in which the $P_{\rm quant}$-command simulated is INPUT, OUTPUT, APPEND, DELETE, IF, or APPLY. The second term covers the simulation of an OBSERVE command: The bound in this case is the product of

[16]To see this, denote by x the amount of probability that $P_{\rm reg}$ has accounted for up to some point. In the worst-case scenario, an amount $1 - p$ (the maximum possible) would have been used on the non-p outcomes (including amount p' on the p'-outcome), leaving just $x - 1 + p$ for the p-outcome. So, in order that there be declared a clear winner at this point, p's amount $(x - 1 + p)$ must exceed p''s amount (p') plus the so-far unallocated probability $(1 - x)$.

our bound ($\mathcal{N}$) on the number of OBSERVE commands in a branch and the difficulty (from (**??**)) required to compute, for each OBSERVE command, the rational number to include in the special section. Combining (**??**) and (**??**), we obtain[17]

$$f_{\rm reg} \leq a^{f_{\rm quant}}, \tag{17}$$

where we have set $a = 4m^{2M_{\rm H}}$.

We conclude: Given any quantum-assisted program, there exists a regular program that computes exactly the same problem, and has difficulty function satisfying (**??**). The benefit in efficiency from the quantum-assist cannot be more than exponential. There is a more elegant, if slightly less informative, way of putting this. For f any difficulty function, denote by $\log f$ the difficulty function obtained by taking the log of f, possibly after adding to f a constant such that it is bounded away from 1. We note that the logs, so defined, of equivalent difficulty functions are equivalent. Then (**??**) implies: $\log f_{\rm reg} \leq f_{\rm quant}$. Of course, these general inequalities are rather coarse. If, in a particular example, a finer inequality is wanted, it usually can be obtained by applying (**??**) directly.

As an example of these ideas, let us return to the Grover construction. Here $m = 2$. Let us assign to each APPLY and OBSERVE command difficulty one; and suppose that none of these operators require a tensor product of more than two H's. Then $M_{\rm op} = 1$ and $M_{\rm H} = 2$. Choose, for $a > 2^{2M_{\rm op}} m^{2M_{\rm H}}$, the value $a = 64$.

Our quantum-assisted program computes this problem with difficulty function $f_{\rm quant}(n) = \sqrt{N}h(n)$, where $N = 2^n$. The inequality (**??**) now implies the existence of a regular program to compute this problem, with difficulty function $f_{\rm reg}(n) \leq (64)^{\sqrt{N}h(n)}$. We can find a much better bound than this. This particular program requires, for given n, just n H's in the tensor product, and it applies to this Hilbert space just $h(n)$ operators. Thus, to simulate a single OBSERVE requires of $P_{\rm reg}$ difficulty $h(n)2^n$. There is a total of n such OBSERVE commands to be executed, and so we obtain $f_{\rm reg}(n) \leq nNh(n)$. Recall, by contrast, that the actual regular program for this construction has an even smaller difficulty function, namely $f_{\rm reg}(n) = Nh(n)$. The extra factor of n in the former reflects the fact that

[17] This equation appears to be nonsensical, in that it is not invariant under replacing $f_{\rm quant}$ by the equivalent difficulty function, $2f_{\rm quant}$. But we broke this invariance in the derivation by assigning to each APPEND and OBSERVE command difficulty one.

our simulation recomputes $\gamma(\mathbf{S})$, from scratch, for each OBSERVation, whereas it is more efficient to carry out these n computations together.

21 Quantum-Assisted Efficiency II

Can quantum-assisted programs offer any gain in efficiency over regular programs? Here is a precise formulation of this question.

Conjecture[18]. There exists a problem π, together with a quantum-assisted program P_{quant} (difficulty function f_{quant}) that computes π, such that: There exists no regular program, P_{reg}, that computes this same problem and whose difficulty function satisfies $f_{\text{reg}} \leq f_{\text{quant}}$.

As far as I am aware, we have neither a proof of nor a counterexample to this conjecture.

Note that, for a proof of the conjecture, one must actually *prove* (not merely suspect) that there exists no regular program at least as efficient as the given quantum-assisted one. What makes this conjecture hard is that we do not currently have good lower limits on the difficulty functions for the regular programs that compute a given problem. A good example is the prime problem: the problem of factoring a given integer n into its prime factors. There is no known method for computing this problem with a regular program whose difficulty function is $\leq (\log n)^s$ for every positive number s. Yet, there are indications ([**?**]) that there exists, for this problem, a quantum-assisted program[19] whose difficulty function does satisfy this condition. Thus, the prime problem is, arguably, a plausible candidate for an example as demanded by the conjecture. Surely, one might think, the most promising

[18]It is tempting to pose a stronger conjecture, asserting the existence of π and P_{quant} such that *every* regular program that computes this problem satisfies $f_{\text{quant}} \leq f_{\text{reg}}$ but not $f_{\text{quant}} \sim f_{\text{reg}}$. But this conjecture is very unlikely to be true (even though, as far as I am aware, we don't have a counterexample). The reason is that, given π and P_{quant}, one can normally design a regular program that, while it may have considerably greater difficulty than P_{quant} for most input strings, is less difficult for an occasional string. We saw examples of this sort of construction in Sect. 12.

[19]However, it is not entirely clear whether there can be designed such a program that meets all the requirements of Sect. 17, in particular, that all operators are fixed initially, independent of the input string.

line to prove the conjecture would be to try to prove that this particular example works, i.e., that there is no regular program that computes the prime problem, and whose difficulty function is $\leq (\log n)^s$ for every $s > 0$. I would like to suggest, however, that this line may not be as promising as it at first appears. There has been an enormous effort, by many talented people over many years, to prove that there is no easy computation of the prime problem. Yet this question remains open — and there is a good chance that it will remain open for some time to come. It may well be, in other words, that it is actually easier to settle the conjecture above than the question of the difficulty of the prime problem. But, of course, the prime problem is also of interest for other reasons.

We remark, below, on a few possible directions for proving (or disproving) this conjecture.

The obvious way to prove the conjecture would be to construct the problem π by a diagonal argument. That is, we would introduce the list of all quantum-assisted programs that compute problems, the list of all regular programs that compute problems, and the list of all input strings. In order to choose what π is on the first string, S_1, we would run a few quantum-assisted programs on this string, as well as a few regular programs, determining, for these runs, what final strings result what the difficulties are. Then, we would select $\pi(S_1)$ so as to to eliminate the low-difficulty regular programs as well as the high-difficulty quantum-assisted programs. Continuing in this way through the list of input strings, we would hope to design a π with a quantum-assisted program that computes π, such that no regular program is at least that efficient. This appears to be a natural line (similar to the Blum ([**?**]) proof of Sect. 10). But, despite its apparent promise, this line has not so far met with success.

Another strategy involves exploiting probabilistic programs. We have seen, in Sect 19, that every probabilistic program may be simulated by a quantum-assisted program with the same difficulty function. Suppose, then, that we were able to find a probabilistic program whose difficulty function is unmatched by any regular program. This would yield immediately a proof of the conjecture. This strategy appears promising, for the language of probabilistic programs is somewhat simpler than that of quantum-assisted programs. Indeed, whereas quantum-assisted programs require additional objects (the Hilbert space H, the state $|\psi_o\rangle$, and some unitary and some projection operators on tensor products of H's), proba-

bilistic programs do not. Whereas quantum-assisted programs require two additional commands (APPLY and OBSERVE), probabilistic programs do not. Whereas quantum-assisted programs require conditions on the function γ on history strings, probabilistic programs do not. The downside of this strategy is that quantum-assisted simulations of probabilistic programs do not appear to exploit what appears to be the key advantage of quantum-assistance — use of entanglements in the tensor product. In any case, as remarked in Sect. 13, there is no known example of a problem and a probabilistic computation of that problem for which one can prove that no regular program is at least as efficient.

Another class of possible examples comes from the Grover construction (Sects. 15-16). Suppose that we could find a suitable example of a needle-in-the-haystack problem. By "suitable", we mean one for which there is a regular program for checking needle candidates; but there exists no regular program can find the needle any more efficiently than merely checking all possible candidates, one at a time. Then, as discussed in Sect. 16, this arrangement might lead to a problem and quantum-assisted program that satisfy the condition of the conjecture. But no such example, apparently, is known.

Another strategy is to try to construct, using the quantum-assisted programming language itself, a problem for which quantum-assisted programs are well-suited but regular programs are not. Consider, for example, the following: Let π accept as input a pair $(P_{\text{quant}}, S_{\text{in}})$, where P_{quant} is a string representing a quantum assisted program, and S_{in} is any string; and let π, on such input, return the string that is determined by running program P_{quant} on input string S_{in}. Quantum-assisted programs are certainly well set up for this π! But, unfortunately, this π is not even a problem, for P_{quant}, applied to S_{in}, need not determine any string at all: There may be a nonzero probability that the program, on this string, will fail to halt altogether; or, even if it does halt, it may do so such that no one output has a probability strictly greater than that of every other possible output string. It would not help to modify this example to read: $\pi(P_{\text{quant}}, S_{\text{in}})$ is the empty string in case P_{quant}, applied to S_{in}, fails to compute any string; and otherwise is whatever string P_{quant} does compute. Now we do indeed have a problem π but, unfortunately, it is not a computable one. Indeed, there exists no program that will even decide whether or not a given regular program and input string results in a halt.

More promising is to focus on the essence of quantum-assisted comput-

ing: the function γ. Fix a Hilbert space H, an initial state $|\psi_o\rangle$, and finite collections of unitary and of projection operators, each acting on some finite tensor product of H's. Recall, from Sect. 18, that a string $\mathbf{S}$ is called a *history* if it represents a finite, ordered, list of unitary or projection operators from the collection above, where for each operator there is specified the particular copies of H on which it is to act; and for each projection operator there is assigned a result ("0" or "1") of an observation via that operator. Fix a history string, $\mathbf{S}$. Then i) assemble a tensor product of H's, consisting of those on which the operators in $\mathbf{S}$ act; ii) consider the state $|\psi_o\rangle \cdots |\psi_o\rangle$ in this tensor product; iii) apply to this state the operators of $\mathbf{S}$, in order, except that, for each projection P that is assigned result "0", apply $I - P$ rather than P; and iv) take the squared-norm of the resulting state. This is the number we denoted $\gamma(\mathbf{S})$ in Sect. 18; and we required there that it be rational-valued. In physical terms, $\gamma(\mathbf{S})$ is the probability that the sequence of operations and observations represented by $\mathbf{S}$ will in fact return the results we have assigned to each of the observations. This function γ contains all the information about quantum mechanics needed to simulate quantum-assisted programs written in this language.

Fix H, $|\psi_o\rangle$, and the collections of unitary and projection operators. Consider the problem γ itself, i.e., the map that assigns to history string $\mathbf{S}$ the (rational) number $\gamma(\mathbf{S})$. This is indeed a problem; and, by virtue of the restrictions we imposed in Sect. 18, it is a computable problem. If we could compute this problem γ by some relatively efficient regular program, then we could simulate each quantum-assisted program by a regular program with the same difficulty function, and thus would conclude that the conjecture is false. But, as we remarked in Sect. 18, there is no obvious efficient method for computing γ with a regular program, and this observation is the essence of the idea that quantum mechanics might make for more efficient computation.

So, let us take, for the problem π of the conjecture, the problem γ itself. It is, as we remarked above, unlikely that this π can be computed efficiently by a regular program. But this example does not seem to work either, for no quantum-assisted program can (at least, not in any obvious way) do any better! Quantum-assisted programs are very good at taking actions in response to probability $\gamma(\mathbf{S})$ (for that is the essence of the OBSERVE command in that language), but they do not seem particularly adept at actually computing the integers that appear in the numerator and the denominator of this fraction.

Here is a more promising way to incorporate this γ into a problem. Let,

for $\mathbf{S}$ any history string, $\pi(\mathbf{S})$ be "yes" if $\gamma(\mathbf{S}) \geq 1/2$, and "no" if $\gamma(\mathbf{S}) < 1/2$. This is, again, a computable problem; and, again, it is plausible that there is no regular program that computes it efficiently. But here is a relatively efficient quantum-assisted program, P_{quant}, for this problem. Let P_{quant}, given a history $\mathbf{S}$, simply perform, physically, the operations represented by $\mathbf{S}$, keeping a record of the results of any OBSERVE commands. Then, P_{quant} compares those actual results with the results already encoded in $\mathbf{S}$, and reports "yes" if they agree, and "no" if they do not. This quantum-assisted program, P_{quant}, computes this problem π, except for one little thing. For any string for which $\gamma(\mathbf{S})$ has exactly the value $1/2$, then this P_{quant} computes nothing (for it will return "yes" or "no", each with probability $1/2$). But this is easily remedied by slightly modifying P_{quant}. First note that, if history string $\mathbf{S}$ contains exactly L projection operations, then $\gamma(\mathbf{S})$ is a fraction with denominator 2^L. Thus, we have only to modify P_{quant} so that, for input string $\mathbf{S}$, it reports "yes" with probability $\gamma(\mathbf{S}) + 1/2^{L+1}$ (rather than just $\gamma(\mathbf{S})$ as before); and "no" otherwise. This is easily accomplished by incorporating into P_{quant} a quantum-generated additional probability of $1/2^{L+1}$ for "yes".

So, we have a quantum-assisted program, P_{quant} that computes the problem π. Note that this P_{quant} is relatively efficient when $\gamma(\mathbf{S})$ is far from the value $1/2$ — say, less than $1/3$ or more than $2/3$. But, when $\gamma(\mathbf{S})$ is close to $1/2$ — and, in particular, when it is precisely $1/2$ — then P_{quant} will be very inefficient, a consequence of the fact that, since the numbers of "yes" and "no" answers will be nearly equal, many runs of P_{quant} will be necessary to determine $\pi(\mathbf{S})$. Indeed, when $\gamma(\mathbf{S})$ is close to $1/2$, P_{quant} may be less efficient than the regular program that computes π. We could (although it is not necessary, in light of the way the conjecture is structured) adjust for this by modifying P_{quant} further. While running its quantum-assisted computation of $\pi(\mathbf{S})$, P_{quant} also carries out the regular computation of $\gamma(\mathbf{S})$. Then, P_{quant} reports whichever method terminates first.

So, we have a problem π, and a quantum-assisted program, P_{quant} that computes π, such that the obvious regular program to compute π is not more efficient that P_{quant}. But this alone does not establish the conjecture: We must *prove* that there exists *no regular program whatever* more efficient that P_{quant}. But — and this will come as no surprise — obtaining such a proof does not seem to be easy.

We remark that there are some pretty tough-looking subproblems of this problem π. For example, let the Hilbert space H be two-dimensional, let

the unitary operators include a "spin-flip" operator (that reverses "up-spin" and "down-spin") and a Toffoli operator (that flips one spin if and only if two others are "up"). Let there be just one projection operator, the "spin-up" projection. Consider history strings that i) create a tensor product of n copies of H (so the Hilbert space of states has dimension 2^n), ii) create some initial state in this Hilbert space, iii) apply to this state m Toffoli operators, acting on various of the factors, and iv) OBSERVE the spin-state in the first H-factor. Now, for many such history strings, $\gamma(\mathbf{S})$ will be close to 1/2; but there will also be many for which $\gamma(\mathbf{S})$ is far from 1/2. For the latter, quantum-assisted program P_{quant} will compute $\pi(\mathbf{S})$ with difficulty $n+m$. It appears that it will be extremely difficult to find a regular program that will be more efficient for these history strings. Thus, it appears plausible that this π and P_{quant} will satisfy the condition of the conjecture.

To summarize, there are a number promising-looking strategies one might employ to try to decide whether the conjecture above is true or false. But none, so far, has panned out. It appears that the question posed by this conjecture is a difficult one.

22 Conclusion

We have discussed here three broad aspects of the theory of computation.

The first of these is the notion of computability — what can, in principle, be computed. This subject is, by any measure, in excellent shape. There is apparently a unique, natural notion of what it means to "compute" something. And we can produce simple examples of problems that are computable and of those that are not.

The second aspect involves the notion of the difficulty of a computation — roughly speaking, the number of steps required to carry it out. We introduce what purports to be "the simplest efficient language", and, by means of it, define what we mean by a "method" to compute a problem, as well as by the difficulty of that method. It has been proven that there exist "very hard problems"; and that there exist problems for which there is no "most efficient" computation. Furthermore, there is a simple, natural way (by merely altering slightly one of the commands of this language) to incorporate probability into the computation process. A probabilistic program returns, for a given input string, an "answer" only probabilistically. Nevertheless, there

is a suitable definition of what it means for such a program to compute a problem; and one can assign, in a natural way, a difficulty function to such a computation. There are at least three open issues in this area. First, while our definition of difficulty is perhaps reasonable, there do exist some technical variants of it, and there is no solid argument that our scheme is "more reasonable" than these alternatives. Can such an argument be found? Second, it turns out that, for virtually every interesting problem, we do not have a good lower bound on the difficulty of the regular computations of that problem. This lack of good lower bounds is arguably *the* outstanding gap in this subject, and an enormous amount of effort has gone into it. And, finally, it is not known whether there exists a problem along with a probabilistic computation of that problem such that no regular computation is equally efficient. It is surprising, to say the least, that such a question should remain open. It would be most interesting to settle it.

The third aspect involves the use of quantum mechanics in the computation process. It turns out that one can introduce a certain, precise computer language, designed to reflect what (and only what) could be done, in the laboratory, using quantum systems. This language introduces "basic quantum systems", with a finite-dimensional Hilbert space, unitary evolution, and certain observables. Computations are carried out using tensor products of copies of these basic systems. One introduces, further, a suitable difficulty function for computations in this language, designed to reflect physical difficulty of an actual computation. This, the language of "quantum assisted computing", allows us to reformulate questions about physical computers into questions about mathematics. Using this formulation, for example, we can prove that, given any computation of a problem using quantum mechanics, there is a computation that does not use quantum mechanics, that is at most exponentially more difficult. A central question is whether or not there exist a problem, together with a quantum-assisted computation of that problem, such that that computation is more efficient than *any* non-quantum computation. Although there certainly are indications that there does exist such a problem, we have today neither a proof nor a counterexample. This, to my mind, is one of the most fascinating questions in the subject of quantum-assisted computing.

Quantum mechanics brings to the computation process two, quite separate, potential advantages. One involves the use of probability. We can write quantum-assisted programs that, essentially, use quantum mechanics only as

a random-number generator; i.e., that merely mimic probabilistic programs. But, even when restricted to this aspect, quantum mechanics has the potential to enhance the computation process: It is an open question whether there exists a probabilistic program (and, therefore, a quantum-assisted program invoking only random-number generation) that cannot be matched, in terms of efficiency, by some regular program. The second involves the use of entanglements, i.e., of the ability, with quantum mechanics, to manipulate large numbers of terms in a tensor product in a single step. It is from this source that the advantages of quantum mechanics — if there are any at all — are likely to be the most dramatic. These two appear to be quite separate effects: They rely on very different features of quantum mechanics. In light of all this, it is strange that the example in which the probability-aspect of quantum mechanics might play a role (Sect. 13), and that in which entanglements might play a role (Sect. 15) are strikingly similar. Is it possible that, on some deeper level, these two aspects are somehow related to each other?

It would be interesting to try to do with classical mechanics what has been done with quantum. That is, we would introduce a precise mathematical language of "classical-mechanics-assisted computing", designed to reflect how classical mechanics might be used in the laboratory. With such a language in hand, we could, for example, ask whether or not there exist a problem and a classical-mechanics-assisted computation of that problem, such that the efficiency of that computation cannot be matched with any regular program. Let us, for example, posit that "classical mechanics" is to be idealized as follows. A system is to be described by a manifold of states (e.g., its phase space) together with a dynamical vector field on that manifold. The integral curves of the dynamical vector field give the evolution of the system through time. We identify certain regions of this manifold as corresponding to the various input strings; and certain other regions to various output strings. This arrangement may now be used to "compute" in the following manner: Given any input string, begin with the system in a state lying in the corresponding input-string region of this manifold. Then evolve (i.e., follow the dynamical vector field) until we arrive at some output-string region. But it turns out, unfortunately, that these particular rules for classical-mechanics-assisted computing are too permissive. It is, for example, not difficult to specify a particular manifold, along with a vector field and such regions, such that the resulting system, under these rules, computes the halting problem. We have merely to encode, into the vector field and the

regions, which programs halt and which do not. Thus, the problem with the framework above is that it does not impose, on the vector field and region-assignments, some suitable requirement to the effect that they be "physically constructible". It is very hard to think of any mathematical condition that could be imposed on the manifold, the vector field, and the regions that would reflect such a requirement.

A similar situation can arise already in the quantum case. Let us idealize quantum mechanics as follows: A quantum system is described by a Hilbert space together with a family of unitary operators, giving the time-evolution, and also with a collection of self-adjoint operators, giving the observables. These observables are to be interpreted as representing the input and output strings. This arrangement could be is used to compute, in a manner similar to that of classical mechanics. Start the system in an appropriate eigenstate of an input-string observable, evolve the system using the unitary operators, and make a final observation via an output-string observables. But, just as with classical mechanics, this arrangement could be used to compute the halting problem. But there is one crucial difference, in this regard, between classical and quantum mechanics. Whereas it appears very difficult, in the case of classical mechanics, to invent new rules that can be imposed to prevent this sort of thing, it is relatively easy to do so in quantum mechanics. We first introduce a very simple quantum system — consisting of a finite-dimensional Hilbert space with a couple of simple unitary and self-adjoint operators thereon. This system is relatively structureless — it is not rich enough to encode, for example, the solution to the halting problem. We now build our quantum-assisted computer by taking (finite) tensor products of copies of this simple system. In other words, we demand that the quantum system we use for our computation be explicitly constructed from these simple building blocks. In this way, we are able to use quantum mechanics to assist in the (complicated) computation of a problem, without the danger of encoding the solution of the problem, right from the beginning, in the quantum system itself. It is difficult to think of any way to do a similar thing for classical mechanics. What, for example, are the analogous building blocks?

It is the tensor product that gives quantum mechanics its potential advantage in efficient computing. The tensor product of n physical systems, each having, say, m states, describes a system whose general state is a superposition of m^n states. We thus can, by relatively simple manipulations (i.e., applying operators to one or two of the n systems), manipulate m^n numbers

(the coefficients in the superposition). Can we find other physical theories that might employ a similar advantage? That is, do we find, in any other physical theories, a "tensor-product-like" construction?

Consider electromagnetism. Suppose that we were capable of manufacturing small boxes, in which there could be stimulated a total of three possible electromagnetic modes. Thus, the electromagnetic states within each box form a 3-dimensional (real) vector space, V. Now take two such boxes, place them side by side, and regard these two as a single system. What is the space of states of this combination? Well, each of the two boxes carries a field, in some state in V, and so the state of the total system is described by simply specifying these two elements of V. That is, the vector space of states is $V \oplus V$, the direct sum of V and V (with dimension $6 = 3 + 3$). Had this instead been $V \otimes V$, the tensor product (with dimension $9 = 3 \times 3$), then we would have the beginnings of a promising theory of electromagnetic-assisted computing.

Tensor products, it appears, do not routinely make an appearance outside of quantum mechanics. Is there some general principle of non-quantum physics that rules out the tensor product, once and for all? The following example may shed some light on this question. Consider a one-dimensional "box", of length L, in which the vector space V of allowed states is that resulting from exciting three modes of a field, given, say, by ($\sin \pi x/L$, $\sin 2\pi x/L$, $\sin 3\pi x/L$). Here, for this example, is a mechanism to realize the tensor product, $V \otimes V$. Consider fields in the square of side L. The corresponding modes are arbitrary linear combinations of products, $\sin a\pi x/L \sin b\pi y/L$, where $a, b = 1, 2, 3$. The vector space of such solutions is indeed the 9-dimensional $V \otimes V$. Similarly, passing to a cubic box, we obtain a space of field states given by $V \otimes V \otimes V$. Here, in other words, is a situation in which we *can*, physically, form tensor-product states. But just this three-fold tensor product is not good enough — we must be able to take arbitrarily large tensor products if this scheme is likely to be useful in computations. Alas, we all too soon run out of dimensions.

It might also be of interest to try to characterize all problems (or, at least, some large class of problems) for which quantum-assisted computation is more efficient than computation without quantum mechanics (assuming, for the moment, that there exist any such problems at all!). Even the question itself must be stated with care, in light of the dependence of computational-efficiency on the method employed. We might like to ask,

for example, whether a given quantum-assisted program for computing a problem is more efficient that the non-quantum program "using the same method". But, unfortunately, we do not at the moment have any notion of "same method" for quantum-assisted and non-quantum-assisted programs. Can we find a nontrivial class of problems for which we can *prove* that, in some suitable sense, for no problem in this class does quantum-assist offer any advantage?

Finally, it might be interesting to understand in some deeper sense how our physical theories interact with the mathematics of computation. Can one, for example, imagine a plausible-looking physical theory within whose framework certain computations can be speeded-up even more dramatically?

23 Appendix A. Formal Systems

Fix, once and for all, some computer language; say, that of Sect 11. Let P be any program. (Recall that have required of a program that it initially accept some input string, and then, if and when that program halts, it produce some output string; with no other input or output.) Further, let A and B be any strings, and n any positive integer. We shall write $H(P, A, n, B)$ to mean "when the program P is run, with input string A, then that program halts precisely on the n^{th} step (i.e., not earlier and not later), then returning as output the string B". They key thing about this $H(P, A, n, B)$ is that it makes an assertion we can check explicitly. That is, we can manually run the program P, on the input string A, and then determine whether or not it continues to run until reaching the n^{th} step, and that it at that point halts, returning output string B.

We shall allow the arguments of $H(\ ,\ ,\ ,\ )$ — i.e., the programs, strings, and integers — to be of two kinds. The first is the *constants*, i.e., arguments that are given explicitly. Thus, the argument "n" could be replaced by the specific integer "274"; or "A" by the explicit string "$x2\&.Vv$". We shall also allow arguments that are *variables*. So, for example, the variable "n" would run over all positive integers; and "P" over all programs in our language.

We now introduce, in terms of this $H(\ ,\ ,\ ,\ )$, *formulae*, according to the following rules:

1. Each $H(P, A, n, B)$, where each of P, A, n, B is either a constant or a variable, is a formula.

2. If F and F' are formulae, then $F \wedge F'$ (read "F and F'"), $F \vee F'$ (read "F or F'") and $\neg F$ (read "not F") are also formulae.

3. If F is any formula, with program-variable P, then $\forall P\ (F)$ (read "for all P, F") and $\exists P\ (F)$ (read "there exists a P such that F") are also formulae. Similarly for "P" replaced by any other choice of program-variable. And similarly also for $\forall$ or $\exists$ preceeding any integer-variable or any string-variable. (It will be clear from context which type of variable is intended.)

These three rules, taken together, result in a large variety of formulae — typically involving many H's, with constants and variables for their arguments, strung together in many different combinations using the symbols $\wedge, \vee, \neg, \forall, \exists$. Of course, the words we have used to describe these symbols reflect how we shall interpret those formulae. But — and this point is crucial — these interpretations are *not* part of the formal system itself.

The idea is that the formulae, formed according to these three rules, are sufficient to express anything might wish to say about the results of running programs. After all, a program does nothing more than accept an input string, run, and then possibly halt returning some output string — and this is precisely what is described by $H(\ ,\ ,\ ,\ )$. Here are a few examples. To express that program P, with input string A, never halts, we would write $\forall n \forall B\ \neg H(P, A, n, B)$. To express the result of using the output from program P as input for program P', we would write $\exists m \exists B\ (H(P, A, m, B) \wedge H(P', B, n, C))$. To express "$F$ implies F'", where F and F' are formulae, we would write $(\neg F) \vee F'$. (This formula is normally abbreviated $F \Rightarrow F'$.) To express "for every string A containing the character z, $\cdots$", we first write a program, Q, that accepts as input a string A, and then halts, reporting "yes" if the string A contains the character "z", and "no" otherwise. Then this idea is written as $\forall A((\exists m\ H(Q, A, m, \text{“yes”})) \Rightarrow \cdots)$. In a similar way, we could quantify, e.g., over all prime integers; or over all strings of symbols that represent formulae. Exercise: How would you write a program that utilizes the output from two other programs?

In fact, the formulae above do much more than merely express everything we might wish to say about programs. Arguably, they express everything we might wish to say within mathematics as a whole. For example, let P be a program which, given input string S, converts that string to an integer n, and then searches systematically for and integer $m > n$ such that m and $m + 2$ are both primes, halting if and when it finds such an m. Then $\forall S \neg(\exists n \exists B\ H(P, S, n, B,))$ is the twin-prime conjecture. But what about an

assertion such as "Every vector space has a basis."? We can hardly write a program that takes as input a vector space, and then searches systematically for a basis for that vector space. This assertion is within set theory. (Thus: A basis is a *subset* of the *set* of vectors of the vector space; and the real numbers (which underlie vector spaces) are defined in terms of Dedekind cuts (*subsets* of the *set* of rationals).) So, to formulate this statement we must turn to a formulation of set theory. One such is the Zermelo-Frankel. Here, assertions about sets are translated into strings of the symbols of this formulation. These symbols include $\in$ ("is an element of"), $\forall a$ ("for all sets a"), $\neg$ ("not"), etc. The statement above, about vector spaces, thus becomes a certain string, A_o, of these symbols. The Zermelo-Frankel formulation further incorporates the notion of a (stylized) "proof". This is again a string of symbols, which represent the "steps" in the proof. And there are rules for which strings represent legitimate proofs of other strings. We needn't be concerned with the details of these rules, except for the following: There is a procedure for deciding whether one string constitutes a proof of another. To put this in other terms, Zermelo-Frankel provides a certain program, P_{ZF}, which accepts as input a string of symbols of this formulation, searches for a "proof" (i.e., one satisfying the rules of this formulation) of that string, and, if and when it finds such a proof, halts, returning "ok". So, the question of whether "Every vector space has a basis." is translated into the following formula: $\exists n \ H(P_{ZF}, A_o, n, \text{"}ok\text{"})$. We may summarize this general point as follows: Mathematics, in the final analysis, consists of nothing more than merely checking things. But computer programs (described via the sentences introduced above) are well-designed for carrying out just such checks.

It is convenient to write $H(P, A, n, \ )$ to stand for $\exists B \ (H(P, A, n, B))$; $H(P, A, \ , B)$ to stand for $\exists n \ (H(P, A, n, B))$; and $H(P, A, \ , \ )$ to stand for $\exists n \exists B \ (H(P, A, n, B))$.

A formula is called a *sentence* if all the variables in that formula have been quantified over, i.e., are subject to either "$\forall$" or "$\exists$". Think of a sentence as a mathematical assertion (about the, possibly complicated, results of running various programs, in various combinations, on various input strings for various numbers of steps.) Note that, for S and S' sentences, so are $S \wedge S'$, $S \vee S'$, and $\neg S$.

Call a sentence *simple* if it contains no variables, i.e., if all the arguments that appear in that sentence are constants. For example, the sentence $H(P_o, A_o, n_o, B_o)$, where P_o, A_o, n_o, and B_o are all constants, is simple. Fur-

thermore, any sentence formed from simple sentences using only $\neg$, $\wedge$ or $\vee$, i.e., using only rule 2, is simple. Clearly, every simple sentence arises in this manner, i.e., a sentence is simple if and only if it does not contain $\forall$ or $\exists$. The key feature of simple sentences is that they can be "checked" explicitly. Thus, we check the sentence $H(P_o, A_o, n_o, B_o)$ by running the program P_o with input string A_o, for n_o steps, and determining whether it halts at that point, with output string B_o. And, clearly, if we can check sentences S and S', then we can also check the sentences $\neg S$, $S \wedge S'$ and $S \vee S'$. To put this in a broader context: There is universal agreement as to what "true" and "false" mean when these terms are applied to simple sentences.

But the situation is considerably less clear-cut for non-simple sentences. For these, we cannot, in the same way, carry out an explicit verification. For example, there is no obvious way to "check", explicitly, that something holds "for every value of n".

The above notwithstanding, we have a shared intuition that *every* sentence about programs — even one involving $\forall$ or $\exists$ — is, in some ultimate sense, either true or false. For example, fix a program and input string. Then it is our intuition that "either that program with that string halts (i.e., that it is true that it halts); or that program runs on forever (i.e., it is false that it halts)". We now wish to understand what this intuition is all about. To this end, we turn to the notion of a proof.

A program $\mathcal{P}$ will be called a *proof-program* provided i) it accepts as input any sentence[20], as defined above; and ii) it then either halts, returning output "proven", or it altogether fails to halt. Think of a $\mathcal{P}$ as "systematically searching for a proof of the given sentence, and, if and when it finds one, announcing the good news." Thus, should this proof-program fail to halt, we interpret this as meaning "there is no proof, at least according to the proof-scheme incorporated into the program $\mathcal{P}$." But, thse interpretative remarks notwithstanding, we are allowing $\mathcal{P}$ to be *any* program with the inputs and outputs described above. We do *not* require that $\mathcal{P}$ actually search for anything. And we certainly do not require that it actually "give the proof" (whatever that means!); nor do we require that $\mathcal{P}$ justify or explain how it has reached its decision (or lack thereof). The announcement itself,

[20]It will be necessary, for later purposes, to allow such a $\mathcal{P}$ to accept as input *any* string (not necessarily a sentence), and, say, fail to halt in the case in which that string is not a sentence.

coming from the program $\mathcal{P}$, *is* the proof.

One can, of course, imagine many candidates for such a program $\mathcal{P}$. One is the program that ignores the input sentence completely, and simply announces "proven" for every sentence; another, the one that again ignores the input sentence, but now always fails to halt. Ideally, we would like to settle, eventually, on some proof-program that closely resembles what we would actually do in writing proofs. We could, if we wished, force some sort of resemblance by imposing various demands on the program $\mathcal{P}$. For example, we might demand that, applied to any simple sentence, the program announces "proven" if and only if the check of that sentence, discussed above, turns out positive. (Exercise: Design a proof-program having this property.) Or, to take a second example, we might demand that if $\mathcal{P}$ assigns "proven" to sentences S and $S \Rightarrow S'$, then it must also assign "proven" to the sentence S'[21]. In any case, we could imagine assembling a list of properties that we would like our proof-program to have, and then searching for a program that has those properties. We might anticipate that, as styles change over time, various new properties of proof-programs might become desirable, and that, as a result, our program $\mathcal{P}$ would be subject to occasional upgrades.

In fact, there is already available a particularly natural candidate for a proof-program — what is called the *Peano program*, $\mathcal{P}_P$. In general terms, this program operates in the following manner: The Peano program first converts the input sentence X (about programs) into an "equivalent sentence" about integers. Then, $\mathcal{P}_P$ searches for a stylized "proof", involving elementary arithmetic, of the latter. Here are some more details.

The program $\mathcal{P}_P$, first of all, has the capability to construct formulae, in much the same way as we constructed formulae above, but now restricted to (non-negative) integers. Thus, in these formulae there appear constant and variable integers; the arithmetic operations, $+$ (addition) and $*$ (multiplication), between integers; and the relations, $=$ (equality) and $<$ (inequality), between integers. Additional formulae result from applying $\neg$ to any formula, inserting $\wedge$ or $\vee$ between formulae, or applying $\forall n$ or $\exists n$ to formulae (where here "n" could be replaced by any integer variable). An example of

[21] We should emphasize at this point that one must be a little cautious about such demands. For example, it would not be a good idea to demand the following: If $\mathcal{P}$ assigns "proven" to the sentence $S \vee S'$, then either it assigns "proven" to S or it does so to S' (or possibly both). As we shall see, it frequently turns out that we know that one of two options must hold, but we don't know which one!

such a formula is: $(1 < n) \wedge \forall m \forall p((m = 1) \vee (p = 1) \vee \neg(n = m * p))$. These formulae, of course, are to be interpreted as assertions about integers. This one, for example, is interpreted as the assertion that "n is prime".

Next, the program $\mathcal{P}_P$ employs some rule, fixed once and for all, to convert strings and programs into integers. For example, it might fix some ordering of the characters that are used, and then use dictionary ordering for the strings and programs composed of those characters. Now consider some input sentence X (about programs). The program $\mathcal{P}_P$ converts this sentence into a corresponding sentence, $\hat{X}$, about integers. The sentence $H(P, A, n, B)$, for example, is converted into the integer-sentence that states that there exists an integer that encodes the n steps the program P would go through, starting from string A, and finally returning B. More complicated sentences, involving many H's and using $\neg, \wedge, \vee, \forall, \exists$, become more complicated sentences involving the integers.

The program $\mathcal{P}_P$ next searches for a "proof" of the integer-assertion $\hat{X}$, in the following manner[22]. First, the program $\mathcal{P}_P$ has access to a certain library of assertions, about integers, called *axioms*. These include various facts about the arithmetic operations, such as $\forall m \forall n \forall p$ $(p * (m + n) = p * m + p * n))$. There are further axioms that represent logical relations on formulae involving integers, such as: For A and B any two such formulae, $(A \wedge B) \Rightarrow A$. There is an infinite number of such axioms, and the program $\mathcal{P}_P$ is capable of deciding whether or not a given string of characters is one of these axioms. A *proof* is a finite, ordered list of formulae, each of which is either an axiom, or a formula B, such that, for some formula A, the formulae A and $A \Rightarrow B$ appeared earlier in the list. The program $\mathcal{P}_P$ searches systematically for a list of symbols that i) satisfies the condition above to be a proof, and ii) is such that the final formula in the list is the original integer-sentence $\hat{X}$. If and when it stumbles upon such a list, the program $\mathcal{P}_P$ halts, returning "proven". If $\mathcal{P}$ never finds such a list, then it never halts.

The Peano program has a number of attractive features. For example, for every simple sentence, $\mathcal{P}_P$ returns "proven" if and only if that sentence checks out. In addition, $\mathcal{P}_P$ has various other "logical properties" that one

[22]The actual Peano system, as usually given, is far more condensed than what we have indicated here. But this version — which anyway we only sketch — is entirely equivalent, and is somewhat easier to think about.

would expect. For example, if $\mathcal{P}_P$ announces "proven" for sentence A, then it does so also for the sentence $A \vee B$, where B is any sentence. And finally, the program $\mathcal{P}_P$ "discovers" standard arguments involving the behavior of programs. For example, let Q be the following program: Given an input string S, Q first converts it to an integer, and then checks, in turn, each integer larger than that one to see if that integer is a prime. If and when it finds such a prime, Q halts returning that prime. Now apply the program $\mathcal{P}_P$ to the sentence $\forall S\ H(Q, S, \ , \ )$, i.e., to the sentence we interpret as saying "for every integer, there is a larger prime". Note that this sentence is *not* simple. We claim that, nevertheless, $\mathcal{P}_P$ will indeed return "proven" in this case: $\mathcal{P}_P$ will find (among others) the standard Euler proof of the infinitude of primes.

While the program $\mathcal{P}_P$ is certainly a natural one, it is not written in stone. One could imagine modifying it in any number of ways. The obvious way would be to add or remove axioms (a strategy that, apparently, hasn't proven very fruitful). Perhaps one could come up with some proposals for more radical alterations. It would also be of interest, I feel, to try to restate the Peano program in a way that avoids the integers entirely. After all, the integers are not playing any substantive role here. There should be some way to rewrite $\mathcal{P}_P$ so that, on receiving its input program, it proceeds by directly, by producing and manipulating only various programs.

Now fix any proof-program $\mathcal{P}$, not necessarily the Peano program. Let X be any sentence. There are two distinct senses in which this program $\mathcal{P}$ can bestow its approval on this sentence. One is represented by the sentence $H(\mathcal{P}, X, \ , \text{"proven"})$; the other, by the sentence $H(\mathcal{P}, X, n_o, \text{"proven"})$, where n_o is some specific (constant) integer. The first we interpret as the assertion that the program $\mathcal{P}$, when applied to the sentence X, will ultimately halt (in some (undetermined) number of steps). We shall abbreviate this sentence $\text{Prf}(X)$. The second sentence goes further, for it specifies the specific number of steps, n_o, that $\mathcal{P}$ will execute before it halts. Note that this second sentence, in contrast to the first, is simple. Thus, we can check $H(\mathcal{P}, X, n_o, \text{"proven"})$ — i.e., there is a universal sense in which it is "true" or "false". We shall write $\vdash X$ to mean that, for some constant integer n_o, a check of $H(\mathcal{P}, X, n_o, \text{"proven"})$ has been carried out, with the positive result. If we wish to make it clear which proof-program is being invoked, we may use subscripts, e.g., write $\text{Prf}_{\mathcal{P}}(X)$ or $\vdash_{\mathcal{P}} X$.

Again, fix some general proof-program $\mathcal{P}$. To what extent can such a

program — maybe the Peano program, or possibly some other — accurately reflect our intuitive ideas as to which sentences are "true"? It turns out that there is a sweeping, yet very simple, answer to this question: There exists no proof-program that even comes close. Indeed, we have the following:

Theorem. Let $\mathcal{P}$ be any proof-program. Then there exists a sentence G such that G is true iff and only if $\mathcal{P}$, acting on G, fails to return "proven".

Proof (virtually the same as that of the halting theorem): Fix, once and for all, some correspondence between strings and programs (for example, that which assigns to each string an integer; and then converts that integer back to a program). For S any string, denote by P_S the corresponding program. We now introduce the following program, $\tilde{P}$: Given any string S, $\tilde{P}$ runs the given proof-program $\mathcal{P}$ on the sentence $\neg H(P_S, S, \ , \ )$. If and when $\mathcal{P}$, applied to this sentence, halts, then $\tilde{P}$ is also to halt, announcing, say, the string "ok". If $\mathcal{P}$, applied to this sentence, fails to halt, then $\tilde{P}$ is also to fail to halt. Now, this $\tilde{P}$ is certainly a program, and so $\tilde{P} = P_{\tilde{S}}$ for some string $\tilde{S}$. Let G stand for the sentence $\neg H(\tilde{P}, \tilde{S}, \ , \ )$.

Suppose, first, that G were false. Then (by definition of G) $H(\tilde{P}, \tilde{S}, \ , \ )$ is true; whence (by definition of H) $\tilde{P}$, run on $\tilde{S}$ halts; whence (by definition of $\tilde{P}$) $\mathcal{P}$, run on $\neg H(P_{\tilde{S}}, \tilde{S}, \ , \ )$, halts; i.e., (by definition of G) $\mathcal{P}$, run on G, returns "proven". Similarly for the supposition that A were true.

This theorem asserts, in short, that: You give me your candidate for a proof-program and I'll construct a sentence on which that program gives the "wrong answer". Note that this holds for *any* proof-program $\mathcal{P}$, no matter how absurd. As an example of the theorem, let $\mathcal{P}$ be the program that always — for any input sentence — halts, announcing "proven". Now, the G, constructed in the proof, asserts a circumstance under which $\mathcal{P}$ fails to halt, and so, for this choice of $\mathcal{P}$, G is false. But this particular program $\mathcal{P}$, applied to this sentence G, halts (as it does for every sentence), announcing "proven". Thus, for this choice of the program $\mathcal{P}$: G is false, but $\mathcal{P}$ announces, for this G, "proven". This outcome is, indeed, consistent with the theorem.

The result above is, in essence, the Gödel theorem. In order to extract from it the more familiar versions of that theorem, we must impose some further conditions on the proof-program $\mathcal{P}$. The following discussion, somewhat informal, will be clarified in Appendix B. Let us now demand:

1. For any sentences X and Y, if $X \Rightarrow Y$ and $\mathrm{Prf}(X)$, then $\mathrm{Prf}(Y)$.
2. For any sentence X, if $\mathrm{Prf}(X)$, then $\mathrm{Prf}(\mathrm{Prf}(X))$.

The first of these is, certainly, a property that we would expect any viable proof-program $\mathcal{P}$ to have. The second is a little more subtle. Consider, as an example, the Peano program, $\mathcal{P}_P$. Given a sentence X, $\mathcal{P}_P$ looks for a string S ("the proof"), bearing a certain relation to X. If and when it finds such a string, $\mathcal{P}_P$ announces "proven". Now fix sentence X, and suppose that $\mathrm{Prf}_{\mathcal{P}_P}(X)$, so there does indeed exist such a string, S. But now we can generate a proof of $\mathrm{Prf}(X)$: This proof simply displays the string S, checks that it bears the required relation to X, and finally notes that $\mathcal{P}_P$, applied to X, will eventually hit upon this particular S. Now convert this argument, just given, into a Peano proof. There results $\mathrm{Prf}(\mathrm{Prf}(X))$. In short, for $\mathcal{P}$ the Peano program $\mathcal{P}_P$, then from $\mathrm{Prf}(X)$ we conclude $\mathrm{Prf}(\mathrm{Prf}(X))$. What we have shown, then, is that the Peano program $\mathcal{P}_P$ satisfies the second condition above. There are various other choices for the program $\mathcal{P}$ that also satisfy this condition. Roughly speaking, these consist, for the most part, of programs that operate by searching for appropriate "proof-strings". Exercise: Give an example of a program $\mathcal{P}$ that fails to satisfy condition 1; of one that fails to satisfy condition 2.

In any case, fix a program $\mathcal{P}$ satisfying these two conditions. Let G be the sentence constructed, for this program $\mathcal{P}$, in the theorem above. Finally, let Con stand for the sentence $\neg(\mathrm{Prf}(G) \wedge \mathrm{Prf}(\neg G))$. This we interpret as "the program $\mathcal{P}$ does not assign 'proven' to both the sentence G and $\neg G$". This is at least one rendition of the notion of that the system whose proofs are governed by $\mathcal{P}$ is consistent. Now consider:

$$\neg G \Rightarrow \mathrm{Prf}(G) \Rightarrow \mathrm{Prf}(G) \wedge \mathrm{Prf}(\mathrm{Prf}(G)) \Rightarrow \mathrm{Prf}(G) \wedge \mathrm{Prf}(\neg G) \Rightarrow \neg\mathrm{Con} \quad (18)$$

The first implication follows from the theorem, the second from condition 2 above, the third from the theorem and condition 1 above and the fourth from the definition of "Con". By a similar (but simpler) argument, the same implications go in the opposite directions. We conclude that G is true if and only if Con is true. But now, from the theorem, we conclude that Con holds if and only if the program $\mathcal{P}$, applied to Con, fails to return "proven". That is, if there is a proof of consistency (as expressed by the sentence Con), then consistency for this system must fail! This is called Gödel's second theorem.

24 Appendix B. A Perspective on Mathematics

There is something unsettling about Appendix A.

The whole idea of mathematics, so we are taught at an early age, is to determine which mathematical assertions are "true". To this end, we devise the notion of a proof: Given a mathematical assertion, the proof of that assertion is supposed to be a convincing argument that that assertion actually *is* true. After a certain amount of socialization, we generally come to agree on what sort of an argument qualifies as a proof. And we all share the conviction that once an assertion is proven, it really is "true". Given any integer n, there really does exist a prime greater than n; and given any continuous function f on $[0,1]$, with $f(0) < 0$ and $f(1) > 0$, there really does exist a number $a \in [0,1]$ with $f(a) = 0$. Our proofs turn out to be remarkably successful, in that they do seem to get us to what is actually true.

Then, at some point, we decide that it would be a good idea to formalize the notion of a proof — to provide precise rules for what does and what does not constitute a proof. Why bother? We all harbor at some level, I think, a nagging suspicion that there is some social component to proofs. To see this, try to explain what a proof is to someone who is not familiar with this concept. It would be reassuring if we could remove this component. We have, after all, managed (via proofs) to place on a more secure footing the issue of which mathematical assertions are true, so why not do the same thing for the proofs themselves? Of even greater concern, there may be disagreements, even at this informal level, as to what constitutes a legitimate proof. Here is an example. Let there be given a computer program that accepts as input any positive integer n, and then halts, with output consisting of two things: A certain mathematical assertion, $A(n)$, and a proof (according to some set of formal rules) of this assertion. Now consider the following assertion "for all n, $A(n)$". We propose the following proof of this: "For each n, we have a proof of $A(n)$, and therefore $A(n)$ is true. So, since $A(n)$ is true for every n, the assertion 'for all n, $A(n)$' is also true." The question is: Is this a legitimate proof? I'm not sure that there would be universal agreement on this point. At the very least, this argument utilizes a novel proof-technique, which might be called into question. (Proofs don't usually make explicit use

of the idea (plausible though it is) that "If there is a proof of something, then that something is true.")

So, for whatever reason, we decide that we shall set up a formal notion of a proof. It turns out, as discussed in Appendix A, that there is a simple, general framework for doing this. In the end, all of mathematics (at least all we have so far!) can be reduced to statements about what certain computer programs will do. It is not difficult to introduce, formally, all possible statements that could be made about the outcomes of all possible computer programs. The role of proofs, then, is to decide, among all these statements, which are "true" and which are "false". We seek a mechanistic notion of proof. That is, we seek a proof program — a program, $\mathcal{P}$, that, when provided as input a sentence about programs, either halts with the announcement "proven", or else fails to halt. And, indeed, it is possible to construct programs — such as the Peano program — that mimic very well the sorts of things we do when we generate acceptable proofs. True, there may be some debate, around the edges, as to exactly what is to go into the final program $\mathcal{P}$. But it appears as though the overall idea — to formalize the notion of a proof — is destined for success.

Now comes the unsettling part: We must confront the (Gödel) theorem of Appendix A. This theorem states, in essence, that, no matter what proof-program $\mathcal{P}$ we settle upon, there will always be some assertion G such that: G is true if and only if program $\mathcal{P}$ fails to bestow upon G the label "proven". Our proof-program $\mathcal{P}$ — no matter how well-chosen — will sometimes give the "wrong answer". In more general terms, there is necessarily a disconnect between where our proofs are *supposed* to be getting us (to what is true) and where our formal proof-program actually *does* get us. And the theorem guarantees that this problem will not be solved by any future improvements in the proof-program $\mathcal{P}$.

A natural reaction to this circumstance would be to question the power of formal systems. The argument would run something like this.

> Choose a relatively conservative proof-program $\mathcal{P}$, for example the Peano program. The proofs generated by this $\mathcal{P}$ will use only steps that are obviously correct (although, admittedly, $\mathcal{P}$ may not exploit some of the more exotic proof-methods). But at least we can be pretty certain that whatever sentences $\mathcal{P}$ designates as "proven" actually *are* true. Now consider, for this $\mathcal{P}$, the sentence

> G given by the theorem. If this G were false, then, according to that theorem, $\mathcal{P}$ would designate G "proven", and therefore (since we believe in $\mathcal{P}$) G must be true. We conclude that in any case this sentence G is true. But now the theorem itself guarantees that this G, which we have already accepted as true, will not be designated by $\mathcal{P}$ as "proven".

In short, the proof-system, represented by $\mathcal{P}$, is inadequate, in that there are true assertions that $\mathcal{P}$ misses. And, of course, this circumstance cannot be corrected by any improvements in $\mathcal{P}$. In other words, there are inherent limitations in the power of formal systems to discover the truth. *We* have insight (however gained) into which mathematical assertions are actually true, but, as we now learn, this insight cannot be reproduced, in its entirety, by any "mechanical process". It is a short step from here to argue, for example, that there can be no mechanistic explanation for the human mind.

Certainly, the Gödel theorem is telling us *something*, and this would seem to be a plausible direction to understand what that something is. But I would like to suggest another direction.

We begin with the following, crucial, point: Our notion of the "truth" of many mathematical assertions has the character of a shared intuition. Consider, for example, the statement "for every integer n, there exists a prime greater than n". We have the standard Euler proof of this assertion. (Consider any prime factor of the integer $(n!+1)$.). But in our mind we go beyond the mere steps of this proof. We think of this statement as "really true" — as sustained on some higher level; as true already before we found this proof; as more than a mere artifice of our particular proof-method. Indeed, we imagine that, if anything, the fact that our proof-method recovers the right result provides support for that method. So, this is our mind set — but where does it come from? It is hard to pin this down. There is, after all, no independent check of this statement: You cannot manually examine every single integer to see if there really is a larger prime. All we have is the proof. Imagine the difficulty you would face if you had to explain your insight to someone who does not share your conviction. How, in short, do you go about deciding which assertions you will designate as "true"? Your answer had better be somewhat vague (e.g., along the lines of "I call 'em like I see 'em."), for otherwise it will be used as the bluprint for a proof-program. And yet this intuition, vague as it is, has a very strong hold on us.

The Gödel theorem, then, represents a clash between what our intuition tells us and what we can recover in concrete terms (i.e., from formal proofs). This isn't the first time in science that there has occurred a clash of this type. For example, it was once our intuition that "position" is an intrinsic property of every particle; that there can be no "levels" of the infinite; that there is in the physical world a universal flowing time. But then along came quantum mechanics; cardinality; and relativity, respectively. All of these earlier clashes were of course resolved long ago. And their resolutions all had the same general character. In each case, we learn to reshape our intuition to fit the external circumstances. This is a struggle, for our intuition is deeply embedded within us. Our initial reaction is to be confused as to where the new boundaries are — as to what of our old intuition we can retain and what we must give up. But, eventually, we adapt: We develop a new intuition, which, ultimately, turns out to be as strong as, and at least as satisfying as, what it replaced.

The proposal, then, is this: We must abandon our intuition that many mathematical assertions are "true" on some deeper level — on a level that goes beyond the proofs of those assertions.

This is more easily said than done. Fix a program P_o, and consider the sentence $\forall S\ H(P_o, S, \ , \ )$, i.e., the sentence we interpret as the statement that P_o halts for every input string. For example, the statement of the infinitude of primes takes this form. Now, we may obtain a proof (according to some set of rules) of this sentence. If so, then we regard it as "proven" (according to those rules). But now we are to think of this as simply a reflection of the rules we happen to be using. We would no longer imagine that it is actually the case, in some more universal sense, that the program P_o *really does* halt for every input string S. Indeed, consider the special case in which program P_o has just three lines: The first accepts an input string; the second returns, say, the string "zz"; and the third halts. It is very hard indeed to resist our intuition that, at least for this program, the sentence $\forall S\ H(P_o, S, 3, \text{“zz”})$ is universally true. But this is a slippery slope: Once one starts truth-labeling things, it is hard to know where to stop. It might be helpful to remember that there are some places in mathematics at which we have already made this transition, at least in part. Examples include the statement that every vector space has a basis, or the continuum hypothesis. Perhaps one could ease into this proposal, e.g., by first adjusting one's language to avoid applying the word "true" to various mathematical assertions, or by imagining that there

appears an annual upgrade of the proof-program $\mathcal{P}$. Eventually, one simply stops thinking in those terms.

This intuition-adjustment, as we shall see, renders the Gödel theorem as relatively benign. But I don't think that that is the central point. This change in our intuition, I suggest, serves to make mathematics simpler and more transparent. The Gödel theorem merely brings this issue (which we would anyway have had to face, eventually) to a head.

But now we face an immediate problem. How, if we avoid designating assertions as "true", are we to get anywhere in mathematics? We may illustrate this dilemma with the following conversation. "Is it true that there is no largest prime?" "Well, we don't say 'true' anymore, but we do have the sentence $\forall S\ H(Q, S, \ , \ )$ (call it A), which expreses this idea." "So, is this sentence true?" "Well, we don't say 'true' anymore, but we do have the sentence $\mathrm{Prf}(A)$, which expresses that your proof-program $\mathcal{P}$ responds, for sentence A, with 'proven'." "But does it actually respond 'proven'?" "Well, there is no 'actually', but we do have the sentence $\mathrm{Prf}(\mathrm{Prf}(A))$ $\cdots$." The danger here, in other words, is that mathematics will degenerate into merely writing down various sentences, with no value judgements at all about them. This is *something*, but it is hardly what we would call "mathematics".

There is a natural solution to this problem. Recall that, in Appendix A, we introduced the notion of a *simple sentence* — a sentence containing only constants but no variables. The simple sentences, we noted there, can be explicitly "checked", in some specific, finite number of steps. Thus, for these there is universal agreement as to whether each sentence is true or false. We now modify our intuition, in the following manner: We allow ourselves a notion of "true" applied to these simple sentences — and these only. Let S be any sentence, n_o any integer, and consider the sentence $H(\mathcal{P}, S, n_o, \text{"proven"})$. The interpretation is that, when the proof-program $\mathcal{P}$ is run on sentence S, it halts, after precisely n_o steps, announcing "proven". This sentence is of course simple. A positive outcome of a check of this sentence is written, as in Appendix A, $\vdash\ S$ (suppressing the n_o). For example, let A be the sentence above (about the infinititude of primes), and let $\mathcal{P}$ be the Peano program. Then the Euler proof gives rise to $\vdash A$. This $\vdash A$ (i.e., a report on the existence of a specific proof, under some specific proof-program), then, is our substitute for "A is true". This symbol provides the one and only link between the abstract formalism on the one hand and the "real world" on the other. Once having decided on a proof-program $\mathcal{P}$, mathematical activity

consists of displaying $\vdash A$ for various sentences A. Think of $\vdash$ as a transitive verb, with object a sentence, meaning "to supply the n_o-value for the $\mathcal{P}$-proof of". We may conjugate this verb: "I have $\vdash$ed A.", or "If you $\vdash A$, then I will $\vdash B$."

It might be argued that this "formalist" path is not going to be very fruitful, because it will interfere with progress in mathematics. I do not understand this argument. I do not see anywhere where mathematics is advanced by having the ability to say (or to believe) that non-simple sentences are, in some universal sense, true. Let our proof-program be, say,the Peano program. Then standard number-theory and basic logical arguments in mathematics are expressed as $\vdash B$, for various sentences B concerning the behavior of computer programs. Furthermore, every theorm, for which we have a proof in Zermelo-Frankel set theory, can be recast in terms of some such $\vdash B$. Note that, even here, we are still using the same proof-program $\mathcal{P}$; and we are still applying $\vdash$ only to sentences about computer programs. So, what is it in mathematics that is missing?

The idea, then, is to confine ourselves to $\vdash$-ing various sentences S. In order to execute this idea, we must first fix, once and for all, some specific proof-program $\mathcal{P}$, which will serve as the basis for this $\vdash$. We choose the Peano program $\mathcal{P}_P$. This is primarily for convenience, for with this choice basic mathematical arguments are indeed reflected by our program $\mathcal{P}$. A further convenience is that, instead of having to say that a simple sentence S "checks out", we can (with this choice of $\mathcal{P}$) merely write $\vdash S$. Thus, we reduce everything to $\vdash$.

As we remarked above, all the standard subjects in mathematics — differential geometry, topological vector spaces, group theory, etc — are already structured, essentially, in the required form. In more detail, we are claiming that mathematics books on these subjects consist entirely of i) checks of various simple sentences (specifically, the $\vdash$-ing of certain (generally, non-simple) sentences), and ii) statements of how people feel about that mathematics. But when we come to the mathematics of sentences, proofs, etc., i.e., to the mathematics described in Appendix A, the situation becomes a little more confusing. We now return briefly to that material.

Consider first the very definition of a formula: A string of symbols formed according to the three rules given in Appendix A. We proceed as follows. Fix, once and for all, a specific program, Q, which we interpret as returning "yes" or "no" according to whether the input string is a formula or

not. Note "we interpret": We cannot actually prove that our Q has this property, for there is no way to turn "Q returns 'yes' or 'no' according to whether the input string is a formula or not" into a sentence. However, we can write the sentence $\forall S(H(Q, S, \ , \text{“yes”}) \vee H(Q, S, \ , \text{“no”}))$, which we interpret as "Q always halts, returning 'yes' or 'no'". Since this sentence is not simple, we cannot assert that it is "true". But (at least, provided our Q was written correctly), we can $\vdash$ it. For F any string, denote the formula $H(Q, F, \ , \text{“yes”})$ by $\mathrm{Form}(F)$. Now consider a statement such as "For F any formula, then $\neg F$ is also a formula." To render this, we must first write a program R, which we interpret as returning whatever string was input to R, but with the symbol "$\neg$" appended on the left. Then we shall have $\vdash \forall F \ (\mathrm{Form}(F) \Rightarrow \forall S(H(R, F, \ , S) \Rightarrow \mathrm{Form}(S)))$. In similar ways, we can build up other properties of "formula forming". (Exercise: Write down a sentence whose interpretation is that "For F and F' any two formulae, $F \wedge F'$ is also a formula." Argue that you can $\vdash$ that sentence.) In a similar way, we may describe other properties of strings. For example, we introduce a formula $\mathrm{Sen}(F)$, interpreted as the statement that string F is a sentence; and $\mathrm{Simp}(F)$, interpreted that F is a simple sentence.

A similar situation obtains for the definition of a proof-program. Denote the formula $\forall S(H(P, S, \ , \text{“proven”}) \vee \neg H(P, S, \ , \ ))$ by $\mathrm{PrfPrg}(P)$. Then, for example, we can $\vdash \mathrm{PrfPrg}(\mathcal{P}_P)$, i.e., we can $\vdash$ a sentence we interpret as the statement that our Peano program really is a proof program. And we also have $\vdash \forall S(\neg \mathrm{Sen}(S) \Rightarrow \neg H(\mathcal{P}_P, S, \ , \ ))$. Note that, although $\mathcal{P}_P$ is a constant program, the sentence $\mathrm{PrfPrg}(\mathcal{P}_P)$ is not simple. Let us introduce a second proof-program, $\mathcal{P}_{ZF}$. This is similar in structure to $\mathcal{P}_P$ — i.e., it takes as input a sentence (also involving computer programs), and it "looks for a proof", also using axioms. But $\mathcal{P}_{ZF}$ utilizes in those proofs also the set-theoretic symbol "$\in$" (and, accordingly, the interpretation of variables as sets). Furthermore, its axioms are those of Zermelo-Frankel (which, effectively, include those of Peano). Of course, $\mathcal{P}_{ZF}$ is also a constant program. Now, if everything has been written correctly, we expect $\vdash \mathrm{PrfPrg}(\mathcal{P}_{ZF})$. Indeed, we further expect that $\vdash \forall A \ (H(\mathcal{P}_P, A, \ , \text{“proven”}) \Rightarrow H(\mathcal{P}_{ZF}, A, \ , \text{“proven”}))$, which we interpret as the statement that "If a sentence is Peano-provable then it is also Zermelo-Frankel-provable". Can we $\vdash \forall A \ (H(\mathcal{P}_{ZF}, A, \ , \text{“proven”}) \Rightarrow H(\mathcal{P}_P, A, \ , \text{“proven”}))$?

Fix any sentence A. Recall that we take $\vdash A$ to mean "we can display an explicit value of the integer n_o such that the sentence $H(\mathcal{P}, A, n_o, \text{‘proven}')$ is

true." Why did we not instead take for the meaning "there exists an integer n such that $H(\mathcal{P}, A, n, \text{`proven}')$ is true"? The problem is that $\exists n\, H(\mathcal{P}, A, n, \text{"proven"})$ is not a simple sentence (i.e., is a not a sentence to which we assign the term "true"), while $H(\mathcal{P}, A, n_o, \text{"proven"})$ is. In short, there are two distinct senses in which we can assert that a given program, with given input, halts: i) we can actually specify the step-number on which this halt occurs (something that can be checked), or ii) we can merely provide a proof (under some proof-scheme) of a sentence that expresses this idea. These are reflected in the two varieties of "proven": $\vdash A$ and $\text{Prf}(A)$. Still more generally, there are two distinct notions of "there exists": The explicit version, in which you actually give the object; and the formal version, in which you construct, and then $\vdash$, a sentence to the effect that the object exists. Note that there is no intermediate version: "There really does exists such an object, but I am not specifying it explicitly".

Fix any sentence A, and suppose that $\vdash A$. Then, we might argue, we must also have $\text{Prf}(A)$. After all, $\vdash A$ provides us with a specific integer, n_o, at which the program $\mathcal{P}$, applied to A, halts; while $\text{Prf}(A)$ merely asserts the existence of such an integer. Is there a theorem to the effect that $\vdash A$ implies $\text{Prf}(A)$? Alas, there cannot be, for we cannot even form a sentence to this effect. The reason is that the symbol $\vdash$ is not among those allowed in our formulae. Rather, this symbol merely provides the link between the formalism and our understanding. What this argument *does* give is a rule of thumb: Given $\vdash A$, you can (if you so choose) $\vdash \text{Prf}(A)$, by first using the n_o provided by the former to argue for $\text{Prf}(A)$ (i.e., to argue that $\mathcal{P}$ applied to A, will eventually find this integer n_o), and then expressing that argument as $\vdash \text{Prf}(A)$. We do have, of course, $\vdash \forall A(H(\mathcal{P}, A, 78, \text{"}\textit{proven}\text{"}) \Rightarrow \text{Prf}(A))$. There is, furthermore, a more general theorem here, namely

$$\vdash \forall A(\text{Prf}(A) \Rightarrow \text{Prf}(\text{Prf}(A))), \tag{19}$$

To carry out the $\vdash$ in (**??**), convert the argument, above, into a Peano proof. Note that there is nothing even close to "if $\text{Prf}(A)$, then $\vdash A$". There are many other "rules of thumb" for $\vdash$ — for example, if $\vdash (A \Rightarrow B)$ and $\vdash A$, then we can $\vdash B$. And, there are many other theorems regarding Prf — for example

$$\vdash \forall A \forall B\ ((\text{Prf}(A \Rightarrow B)) \Rightarrow (\text{Prf}(A) \Rightarrow \text{Prf}B)). \tag{20}$$

(Note, by contrast, that there are no theorems on $\vdash$, and no rules of thumb on Prf.) Exercise: At the beginning of this Appendix, we considered the

example of a program that, for each positive integer n, produces a sentence $A(n)$ and a proof of that sentence; and asked whether this circumstance represents a proof of $\forall n\ A(n)$. Reformulate all this in terms of sentences and $\vdash$.

The paragraph above described some properties of $\vdash$ and Prf. What is the link between "what is proven" and "what is true"? There certainly exist specific sentences A_o such that $\vdash (\mathrm{Prf}(A_o) \Rightarrow A_o)$. An example is the A_o given by the sentence $B_o \vee \neg B_o$, where B_o is any sentence. That is, there do exist sentences A_o such that we can $\vdash$ a sentence to the effect that "If A_o is proven, then A_o is true." Is this the case for all sentences A_o? That is, can we

$$\vdash \forall A(\mathrm{Sen}(A) \Rightarrow (\mathrm{Prf}(A) \Rightarrow A))? \tag{21}$$

It turns out that we cannot, but for a curious reason: What follows the $\vdash$ in (**??**) is not even a sentence! To see this, first note that the "$\forall A$" permits A to be an arbitrary string. This is not a problem for "Sen(A)", for this is a formula for every string A. But it is a problem for the A that appears at the end. (True, that A is written already after "Sen(A) $\Rightarrow$", but this doesn't mean that "A" thereafter has suddenly *become* a sentence.) To put this in more concrete terms, if $\forall A$(something) is a formula, then certainly "something" must be a formula for every choice of the string A. So, try, in (**??**), the choice $A =$ "$8Z;;)a$". Indeed, in still more concrete terms we have

$$\vdash \neg\mathrm{Sen}(\forall A(\mathrm{Sen}(A) \Rightarrow (\mathrm{Prf}(A) \Rightarrow A))). \tag{22}$$

where "Sen" is the formula, introduced earlier, that we interpret as testing for sentence-status. Of course, if someone wrote down a constant sentence, A_o, and also managed to $\vdash (\mathrm{Prf}(A_o) \wedge \neg A_o)$, that would be of interest. Similar remarks apply to $A \Rightarrow \mathrm{Prf}(A)$.

We next turn to the Gödel theorem, as discussed in Appendix A. To simplify the discussion, let us fix the proof-program $\mathcal{P}$, mentioned in that Theorem, to be the Peano program — the same one we are using for $\vdash$. (We shall remark later on what happens when we allow other choices.) With this choice, that theorem becomes "There exists a sentence G such that G is true if and only if the proof-program $\mathcal{P}$, applied to G, fails to return 'proven'." This statement is, to say the least, confusing. What does "G is true" mean? We went to this enormous effort — introducing the notion of a proof-program, selecting a good one as our standard, etc — precisely in order

to formalize proofs. We did all this so that we wouldn't have to contend with informal, intuitive proofs, and with their inevitable fallout: a corresponding informal, intuitive notion of "true". Instead of saying that something is true, we are to say that our proof-program, $\mathcal{P}$, applied to that something, returns "proven". So, let us make this substitution in the statement above. The result is: "There exists a sentence G such that the proof-program $\mathcal{P}$, applied to G, returns 'proven' if and only if the proof program $\mathcal{P}$, applied to G, fails to return 'proven'." Surely, this can't be what the theorem is saying! Fortunately, we now have at our disposal a direct way of deciding exactly what the theorem *is* saying: We go through its (informal) proof, and determine what sentence that proof is the $\vdash$ of. Let us denote by G_o the sentence given in that proof, i.e., (in that notation) $\neg H(\tilde{P}, \tilde{S},\ ,\ )$. Then, translating the steps in the proof of the theorem, we find,

$$\vdash ((G_o \Rightarrow \neg \mathrm{Prf}(G_o)) \wedge (\neg G_o \Rightarrow \mathrm{Prf}(G_o))) \tag{23}$$

This, then, is Gödel's theorem.

Note that Gödel's theorem, now, simply provides $\vdash$ for some, perfectly respectable, sentence. There is, as far as I can see, nothing particularly unsettling about (**??**). We no longer have to consider, and contend with the consequences of, "What if G_o is true?" or "What if G_o is false?", for G_o is a non-simple sentence. Eqn. (**??**) does not seem to suggest any "limitations on the power of the formal method", nor indeed, on much of anything. The argument to this effect, given at the beginning of this appendix, now reduces benignly to $\vdash (\mathrm{Prf}(G_o) \Rightarrow G_o) \Rightarrow (G_o \wedge \neg \mathrm{Prf}(G_o))$. (This follows from (**??**) and $\vdash (G_o \vee \neg G_o)$.) These are now merely sentences that (like so many others) woke up one morning to find themselves $\vdash$ed.

We next consider briefly the theorem in its original form (for the case of an arbitrary proof-program, $\mathcal{P}$). Now, there are two proof-programs in play: the variable program $\mathcal{P}$ of the theorem, and the Peano program, $\mathcal{P}_P$, that we use for $\vdash$. Now (**??**) is replaced by

$$\vdash_P \forall \mathcal{P}(\mathrm{PrfPrg}(\mathcal{P}) \Rightarrow ((G(\mathcal{P}) \Rightarrow \neg \mathrm{Prf}_{\mathcal{P}}(G(\mathcal{P})) \wedge (\neg G(\mathcal{P}) \Rightarrow \mathrm{Prf}_{\mathcal{P}}(G(\mathcal{P}))))). \tag{24}$$

Note, first, that the $\vdash$ in (**??**) is with respect to the Peano program. This is necessary, for we need a program here rich enough to encompass the steps of the proof in Appendix A. Next, there appears in (**??**) $\forall \mathcal{P}(\mathrm{PrfPrg}(\mathcal{P}) \Rightarrow$. That is, what follows is to hold for *every* proof program (and not just for one

particular one, as in (**??**)). Note that the sentence G_o, a specific sentence, in (**??**), is now replaced by a formula, $G(\mathcal{P})$, in (**??**). In more detail, this formula is (in the notation of the proof) $\neg H(\tilde{P}, \tilde{S}, \ , \ )$. (It depends on $\mathcal{P}$ through the dependence of $\tilde{P}$ and $\tilde{S}$ on $\mathcal{P}$). And finally, we make it explicit that "Prf" is with respect to the (variable) program $\mathcal{P}$. The expression (**??**), then, carries the full content of the theorem in Appendix A. It all seems very straightforward.

We turn next to the issue of consistency of the Peano program, as discussed at the end of Appendix A. Let us again fix the program $\mathcal{P}$ in the statement of the Theorem to be the Peano program (i.e., the same one we are already using for $\vdash$). And, again, let G_o be the sentence given, in the notation of the proof, by $\neg H(\tilde{P}, \tilde{S}, \ , \ )$. Finally, let us denote by Con the sentence, $\neg(\mathrm{Prf}(G_o) \wedge \mathrm{Prf}(\neg G_o))$, as in that appendix.

Consider now the two conditions listed in that discussion. We have already translated condition 2 into a theorem (about $\mathcal{P}$), namely (**??**). But what about condition 1? We cannot, as we have already seen, form sentences in the manner of $\forall X \forall Y(((X \Rightarrow Y) \wedge \mathrm{Prf}(X)) \Rightarrow \mathrm{Prf}(Y))$, for "$\forall X$" allows X to be an arbitrary string, which wouldn't make sense in this expression. We first note that condition 1 was actually used at only one place in the argument — in the third step of (**??**). But this step requires only $\mathrm{Prf}(\mathrm{Prf}(G_o)) \Rightarrow \mathrm{Prf}(\neg G_o)$ (and its converse). But this already follows, from (**??**) and (**??**).

Thus, the argument at the end of Appendix A can be restructured as a Peano proof, i.e., we obtain $\vdash ((G_o \Rightarrow \mathrm{Con}) \wedge (\mathrm{Con} \Rightarrow G_o))$. From this, together with (**??**), there follows

$$\vdash (\mathrm{Prf}(\mathrm{Con}) \Rightarrow \neg\mathrm{Con}) \wedge (\neg\mathrm{Con} \Rightarrow \mathrm{Prf}(\mathrm{Con})). \tag{25}$$

This, then, is Gödel's second theorem. What can we make of it? There has been considerable speculation, over the years, as to whether or not "the Peano program is consistent". This is a conversation we need not join. From the present perspective, there is no promise of an "ultimate answer" to this question. True we can $\vdash (\mathrm{Con} \vee \neg\mathrm{Con})$. (This is a general property of the Peano proof-program.) But at the moment we cannot $\vdash \mathrm{Con}$, and we cannot $\vdash \neg\mathrm{Con}$. This is all there is to say. There is no sentence we can interpret as the statement that, ultimately, either $\vdash \mathrm{Con}$ or else $\neg \vdash \mathrm{Con}$. Indeed, there is no meaning we can attach to "$\neg \vdash \cdots$" ("I am *not* in the

process of presenting to you the n_o-value for the proof of $\cdots$"?), nor, indeed, to "$\forall A \vdash \cdots$". Mathematics, from this perspective, consists only of $\vdash$-ing sentences (and, in some cases, of failing to $\vdash$ them).

It is my contention that in thinking in this way we lose nothing of the structure of mathematics itself. After all, as we remarked earlier, mathematics per se consists solely of checking simple sentences (specifically, of $\vdash$-ing various sentences).

Yet certainly mathematics, in its entirety, is much more than this. The real substance of mathematics, of course, consists of figuring out which sentences we shall try to $\vdash$, and of actually carrying this out. (Simply running the program $\mathcal{P}$ on randomly selected sentences, for example, is not a very fruitful strategy!) Indeed, how we feel about what we prove is, arguably, more interesting than the formal sentences or proofs themselves. For example, it is hard to discern, within the formal structure of mathematics, how or why mathematics plays such an important role in physics (nor, conversely, why physics has the impact it does on mathematics). Recognizing what already is — that the core of mathematics consists of checking simple sentences — does not diminish the role of insight. Even on learning, from quantum mechanics, that there is no such thing as the position of a particle, we still have "positional thoughts" about quantum systems. It is just that our thoughts are of a different sort.

A key difference between "informal" and "formal" mathematics — and a key source of confusion — is the relation between "something is proven" and "something is true". Indeed, as we remarked earlier, there is not even a sentence whose interpretation is "For every sentence A, $\mathrm{Prf}(A) \Rightarrow A$." One might imagine that, nevertheless, there might be some, more specific, theorems along these lines. We now return briefly to this issue. Can we, for example, $\vdash \mathrm{Prf}(A_o) \Rightarrow A_o$ for some large variety of specific (constant) sentences A_o? Such a program does not look very promising. Consider, for example, the sentence G_o of the Theorem. We claim $\vdash (\mathrm{Prf}(G_o) \Rightarrow G_o) \Rightarrow \mathrm{Con}$. (This follows from two earlier results: $\vdash (\mathrm{Prf}(G_o) \Rightarrow G_o) \Rightarrow G_o \Rightarrow \mathrm{Con}$.) This we interpret as "If $\mathrm{Prf}(G_o) \Rightarrow G_o$, then the Peano proof-program is consistent." But it now follows directly, from this, that $\vdash \mathrm{Prf}(\mathrm{Prf}(G_o) \Rightarrow G_o) \Rightarrow \neg\mathrm{Con}$. In other words "But if you can prove $\mathrm{Prf}(G_o) \Rightarrow G_o$ then the Peano proof program is not consistent." (Exercise: $\vdash ((\mathrm{Prf}(X) \Rightarrow X) \Rightarrow \neg\mathrm{Prf}(X))$, where X denotes $\mathrm{Prf}(G_o) \Rightarrow G_o$.) Perhaps it is easier just to dwell on the sentences themselves, rather than on their "interpretations".

Finally, we remark on one further issue: There is, as it turns out, a certain ambiguity in the notion of $\vdash$. Fix a (constant) sentence S_o, as well as a choice of a proof-program $\mathcal{P}$. Consider the sentence $H(\mathcal{P}, S_o, 2311, \text{“proven”})$. This sentence is certainly simple. We can check it, and if that check works out we are clearly justified in writing $\vdash S_o$. Now consider instead the sentence $H(\mathcal{P}, S_o, 945!, \text{“proven”})$. Again, we would expect to be able to check this sentence, and, again, to write $\vdash S_o$ if that check works out. But this sentence has a slightly different character from its predecessor. Here, we aren't exactly “given” the value of n_o. Rather, we are told that, to compute n_o we must multiply $1 * 2 * 3 * \cdots * 945$. That is, we are given the *instructions* to find n_o. But if we accept that instructions for n_o suffice, then suppose that we are given instead a program that will generate n_o for us. Is this acceptable? Presumably, it is, provided that program actually halts. But what does “actually halts” mean? That it be “obvious” that it halts? That I give you my promise that it will halt? That the Peano program applied to the corresponding sentence, return “proven”? The issue, then, is what, exactly, it means to give a number “explicitly”. Clearly, there is a slippery slope here, from writing out the actual digits of n_o to providing some more elusive specification. In order to make sense of $\vdash$ one has to draw a line somewhere — one has to establish what degree of specificity is necessary in order to grant $\vdash$ status. But where, exactly, is this line to be drawn? Here, just as an example, is a possible position on this issue. “I maintain my own private list of sentences that I have personally $\vdash$-ed. If, for instance, instructions are given for n_o, then I personally follow those instructions, recover the value of n_o, and then check the (simple) formula. I do not add sentences to my list based solely on any claims by others.” There may be more attractive positions.

25 Appendix C. Blum Example Revisited

In Chapter 10, we discussed an example, due to Blum, of a computable problem such that there is no “fastest” program that computes that problem. In more detail, the problem given there had the following property: Given any program that computes that problem, there is another that computes the same problem, but does so, say, exponentially faster than the first. In this Appendix, we look at that example again, in light of the material in

Appendix A. We begin with a few general remarks on how formal proofs can be applied to computable problems.

Fix, once and for all, a computable problem π_o (so π_o is a map from input strings to output strings; and there exists some program that computes this map). We are interested in studying various programs that compute this problem. It is convenient first to describe informally what we have in mind.

We first introduce a subroutine, $\mathcal{R}$, which systematically searches for pairs, (P, A), where P is a program, and A is a proof (in some formal system – more on this later) that P computes the problem π_o. It carries out this search in the following manner. Program $\mathcal{R}$ fixes an ordering (say, dictionary ordering) for all pair of strings. It then systematically goes through those pairs, one at a time, in a pattern such that every pair is, eventually tested. For each pair, it determines whether the first of the two strings is actually a program (as opposed, say, to some gibberish of symbols), and then whether the second string is actually a proof that that program computes the problem π_o. If both these tests are positive, the subroutine $\mathcal{R}$ reports that program.

In this way, $\mathcal{R}$ generates a sequence, $P_1, P_2, \cdots$, of programs, together with, for each n, a proof that P_n computes the problem π_o. Note that, by virtue of this construction, we are guaranteed that *every* program that can be proven to compute π_o will be included in our list. In fact, every such program will be included many times (for we expect that there will be many proofs). Furthermore, for each program among the P_n there will be included in this list various trivial variants of that program, for example, a variant that adds an irrelevant line. The program $\mathcal{R}$, after all, is not equipped to filter out such duplication.

Next, we introduce the following program, Q. This Q first accepts as input some string, S. Then Q begins by running the subroutine $\mathcal{R}$ above. Eventually, $\mathcal{R}$ will find some program, P_1, together with a proof that P_1 computes the problem π_o. At this point the program Q splits its efforts: It allocates half its steps to running P_1 with input string S and the other half to continuing to run $\mathcal{R}$. Eventually, $\mathcal{R}$ will find a second program, P_2, together with a proof that P_2 computes the problem π_o. At this point, program Q again splits its efforts: It continues to allocate half its steps to P_1, but now allocates one-quarter of its steps to running P_2 with input string S. With the one-quarter of its steps that remain, Q continues to run $\mathcal{R}$. Program Q continues in this way: Each time $\mathcal{R}$ finds a new program, P_n, together with a proof that P_n computes the problem π_o, program Q devotes half the steps

it had been using on $\mathcal{R}$ to running program P_n with input string S, keeping the allocations to the earlier P_i intact.

As time goes by, then, program Q will be running a number of programs, $P_1, P_2, \cdots$, all on the input string S. At some point, one of these programs, say P_n, may halt, returning some output string. If and when this happens, program Q itself halts, returning that same string.

So, given any computable problem π_o we construct in this way a program Q. We now claim that it has two properties.

First, this Q computes the problem π_o. To see this, run Q on some input string S. Then, Q, by construction, will only halt when it finds that one of the P_n has halted on string S; and then Q will return as its output whatever P_n returned. But this P_n was generated by the subroutine $\mathcal{R}$, and so there was produced (by $\mathcal{R}$) a proof that P_n computes the problem π_o. Thus, with input string S, program P_n, and therefore program Q, returns $\pi_o(S)$. That is, program Q computes problem π_o.

For the second property, consider any program, P, such that there is a proof that P computes the problem π_o. Then, by our construction of $\mathcal{R}$, this P must have occured among the programs, $P_1, P_2, \cdots$, that $\mathcal{R}$ produced. Say, it was program P_n. Now run program Q on some string S. Then Q will (since Q is running $\mathcal{R}$) eventually run across this program P_n. When it does so, it will allocate a portion $1/2^n$ of its steps to the running of P_n (i.e., to the running of P itself) on the string S. So, except for a constant (the number of steps Q used prior to $\mathcal{R}$'s finding P_n), Q, with input string S, will take no more than 2^n times as many steps as does P, with input string S. (It may be much less: For example, Q may halt even before it gets to P_n.) Note that this constant, and this factor 2^n, are independent of the string S. But difficulty functions make sense only up to addition of constants and multiplication by constant factors. We conclude: The difficulty function of Q is $<$ the difficulty function of P.

To summarize: Given any computable problem π_o, there exists a program Q that computes that problem, and is such that Q is at least as efficient as any program P for which there is a proof that P computes that problem.

Clearly, this construction is, at least naively, in conflict with the Blum example. We are accustomed, in everyday mathematics, to blurring the distinction between "something is proven" and "something is true". If we allow ourselves that luxury here, then we seem to arrive at a contradiction. Our purpose is to understand this matter.

The first step is to reformulate the construction above within the framework of Appendix A.

To begin with, we cannot refer directly to "problem π_o", for this is a mapping from a certain set (that of all strings) to itself; while formal systems are not adapted to handling sets (at least, not without substantial additional infrastructure). But it is easy to get around this difficulty. For P any program, denote by $\mathrm{Tot}(P)$ the formula $\forall A\ (H(P, A, \ , \ ))$. We interpret $\mathrm{Tot}(P)$ as the statement that "for every input string P halts", i.e., as the statement that P computes *some* problem. We can now replace the problem π_o by some fixed (i.e., constant) program P_o, such that $\vdash \mathrm{Tot}(P_o)$, i.e., we can replace the problem by a program that computes it. Now, for P and P' any two programs, denote by $(P \approx P')$ the formula $\forall A \forall B((H(P, A, \ , B) \Rightarrow H(P', A, \ , B)) \wedge (H(P', A, \ , B) \Rightarrow H(P, A, \ , B)))$. We interpret this as the statement that programs P and P' have precisely the same behavior, i.e., that, for any given input, they produce exactly the same output (or lack thereof). For example, we have $\vdash \forall P \forall P'\ ((\mathrm{Tot}(P) \wedge (P \approx P')) \Rightarrow \mathrm{Tot}(P'))$. The intuitive idea that "program P computes the problem defined by P_o" is expressed formally by $P \approx P_o$.

Now fix program P_o, and let $\vdash \mathrm{Tot}(P_o)$. We begin with the subroutine $\mathcal{R}$. This subroutine searches for programs P, together with a "proof" that P computes the same problem as does P_o. The proofs we have in mind now, of course, are formal proofs, as discussed in Appendix A. Thus, we must fix a proof-program $\mathcal{P}$, that will serve as our standard for acceptable proofs. For this purpose, we choose the Peano program, as described in Appendix A. (This choice merely avoids unnecessary complication. Others are certainly possible.) The subroutine $\mathcal{R}$ works as follows. It both searches for programs P and applies the proof-program $\mathcal{P}$ to the sentences $(P \approx P_o)$. It does so in an interlaced way, i.e., it spends some time searching for new programs, then some time running $\mathcal{P}$ on various of these sentences for programs it has already identified, then back to searching for more programs, etc. The first program P for which it finds that $\mathcal{P}$, applied to $(P \approx P_o)$, returns "proven", subroutine $\mathcal{R}$ reports as P_1; the next as P_2; etc. In this way, $\mathcal{R}$ generates a sequence of programs $P_1, P_2, \cdots$ such that, for each, $\mathcal{R}$ has determined $\mathrm{Prf}(P_n \approx P_o)$.

We now turn to the program Q. As described earlier, this program, given input string S, divides its time between running $\mathcal{R}$ (to search for programs P_n, such that $\mathcal{P}$, applied to the sentence $(P_n \approx P_o)$, returns "proven") and

running, on input string S, the programs P_n so selected. If and when one of these P_n halts, then Q itself halts, returning whatever P_n returned.

So, this is our program Q. We wish to determine its properties. The first is that $\vdash \mathrm{Tot}(Q)$, i.e., that that Q computes *some* problem. This follows from the fact that, among the programs $\mathcal{R}$ finds will be the program P_o itself (which, in turn follows from $\vdash (P_o \approx P_o)$). Next, we would like to $\vdash (Q \approx P_o)$, i.e., to prove that the problem that Q computes is in fact the same as the problem that P_o computes. Surprisingly enough, there is no obvious way to do this. To see why, we first establish a weaker result — that $(Q \approx P_o)$ holds under a further hypothesis. Denote by X the sentence

$$\forall P \ (\mathrm{Prf}(P \approx P_o) \Rightarrow (P \approx P_o)). \tag{26}$$

The interpretation of (**??**) is: "If you can prove that a program computes the problem of P_o, then it does indeed compute that problem." We now claim

$$\vdash X \Rightarrow (Q \approx P_o)). \tag{27}$$

To establish (**??**), we proceed as follows. The key features of the construction of the program Q can be summarized by two sentences:

$$\vdash \ \forall S \forall B \ (H(Q, S, \ , B) \Rightarrow \exists P(H(P, S, \ , B) \wedge \mathrm{Prf}(P \approx P_o))), \tag{28}$$

and

$$\vdash \ \forall S \forall B((\forall P(\mathrm{Prf}(P \approx P_o) \Rightarrow H(P, S, \ , B))) \Rightarrow H(Q, S, \ , B)). \tag{29}$$

The interpretation of (**??**) is: "If Q, with input S, returns output string B, then there exists a program P such that i) P with input S, also returns B; and ii) there exists a proof that P and P_o have the same outputs." The interpretation of (**??**) is: "If every program P for which there exists a proof that $P \approx P_o$ has the property that, with input S, this P returns output B, then program Q also has the property that, with input S, it returns B." These two interpretations do, indeed, reflect the way we constructed the program Q. Now, it follows from (**??**) that

$$\vdash X \Rightarrow (\forall S \forall B \ (H(Q, S, \ , B) \Rightarrow H(P_o, S, \ , B))). \tag{30}$$

To see this, first note that, from X and (**??**) we can achieve (**??**), but now with the "Prf" on the right omitted. But from this, and the definition of

$\approx$, we achieve $\forall S \forall B\ (H(Q, S,\ , B) \Rightarrow H(P_o, S,\ B))$. This yields (**??**). In a similar way, from (**??**) we obtain

$$\vdash X \Rightarrow (\forall S \forall B\ (H(P_o, S,\ , B) \Rightarrow H(Q, S,\ , B))). \tag{31}$$

Combining (**??**) and (**??**), there follows (**??**).

The final result, (**??**), is what we might have expected. The program Q selects (through $\mathcal{R}$) programs P based on $\mathrm{Prf}(P \approx P_o)$. But, in order to $\vdash (Q \approx P_o)$, we need that those programs actually *satisfy* $(P \approx P_o)$. So, we must get from $\mathrm{Prf}(P \approx P_o)$ to $(P \approx P_o)$, and that is what the sentence X of (**??**) does for us. Why didn't we write the subroutine $\mathcal{R}$, right at the beginning, so that it selects programs P based on $(P \approx P_o)$ rather than on $\mathrm{Prf}(P \approx P_o)$? The reason is that there is no obvious way to do this: Whereas programs can easily search for a proof of a given sentence, they are unable to search for whether that sentence actually "holds"[23].

In any case, we would, via (**??**), be able to $\vdash (Q \approx P_o)$ if only we could $\vdash X$, where X is the sentence given by (**??**). Can we do this? The sentence X looks very plausible: Intuitively, We would certainly expect that, if you can prove something, then that something must be true. This is precisely the issue we discussed in Appendix B — and that discussion suggests that $\vdash X$ will be far from easy! Indeed, even for the case of the simplest programs P_o — e.g., the one that ignores the input string and always returns, say, the string "zz" — it is not apparent how to $\vdash X$.

For P and P' any programs, let us write $(P < P')$ for the sentence whose interpretation is that $\mathrm{Tot}(P)$ and $\mathrm{Tot}(P')$, and that the difficulty function of P is $<$ (in the sense discussed in Chapter 9) that of P'. (Exercise: Describe this sentence in more detail. You will have to write a program or two.) Note that we do not require in this sentence that P and P' compute the same problem. Similarly for $P \ll P'$. Then we have, for example $\vdash \forall P \forall P'\ ((P \ll P') \Rightarrow (P < P'))$. From the construction of Q (specifically, that Q finds, via $\mathcal{R}$, programs P with $\mathrm{Prf}(P \approx P_o)$ and then incorporates those programs into the running of Q), we have

$$\vdash (\forall P)((\mathrm{Tot}(P) \wedge \mathrm{Prf}(P \approx P_o)) \Rightarrow (Q < P)). \tag{32}$$

Eqn. (**??**) is the formal statement that "program Q is at least as efficient as any program that can be proven to compute the problem of P_o".

[23]The discussion above suggests that, even though we are unable to $\vdash (Q \approx P_o)$, it might be possible to $\vdash \mathrm{Prf}(Q \approx P_o)$ (with no further hypothesis). Is it?

To summarize, given any program P_o, with $\vdash \text{Tot}(P_o)$, we construct a certain program Q, with $\vdash \text{Tot}(Q)$. This Q is so constructed that it has two key features, represented by (**??**) and (**??**), where X is the sentence given by (**??**).

We would now like to compare this construction — which provides, given any computable problem, a program that is at least as fast as any program that can be proven to compute that problem — with the example of Chapter 10 — a problem for which there is no fastest program. To this end, we now return to that example.

In Chapter 10, there was written down a certain program, P_o, structured as follows. First, P_o employs a subroutine, which finds, in sequence, all possible programs, $T_1, T_2, \cdots$. Given an input string S, P_o first converts S into an integer, say k, then uses this subroutine to find the first k of these programs, $T_1, T_2, \cdots T_k$, and finally runs each of these k programs on input string S. These programs are run for certain numbers of steps (determined using the function h defined in Chapter 10). From the results of these runs, program P_o determines what its output is to be. This P_o, so constructed, is indeed a program, and we have immediately from this construction $\vdash \text{Tot}(P_o)$. Thus, P_o defines a problem, π_o, and this is the problem of interest.

The key result of Chapter 10, is that the program P_o there constructed has the property that there is no "fastest" program that computes that problem, i.e.,

$$\vdash \forall P((P \approx P_o) \Rightarrow \exists P'((P' \approx P_o) \wedge (P' \ll P_o)). \tag{33}$$

The argument for this, in more detail, is the following. We introduce a certain sequence of programs, $P_1, P_2, \cdots$. Each P_n operates in the same manner as P_o, but with one difference. Whereas P_o uses the entire sequence $T_1, T_2, \cdots$ of programs, P_n uses only $T_{n+1}, T_{n+2}, \cdots$. As discussed in Chapter 10, three things follow from the construction of these P_n:

1. $\vdash \forall n$ Tot(P_n), i.e., each P_n computes some problem.

2. Each P_n produces the same output as P_o, except possibly for a finite number of input strings.

3. $\vdash \forall n$ $((T_n \approx P_o) \Rightarrow (P_n \ll T_n))$, i.e., if T_n computes the problem of P_o, then P_n is much faster than T_n.

From these three properties, we establish (**??**), as follows. Let P be any program that computes the problem of P_o. Then P must have been one of the T_i, say T_n. Now consider the corresponding program P_n. By the third

property above, P_n is much faster than T_n, while, by the first two properties, P_n computes *some* problem, and that problem differs from that of P_o on only a finite number of input strings. So, consider all programs whose output differs from P_n only only a finite number of strings. One of these, call it P', must compute the same problem as P_o. Now (**??**) follows.

Let P_o be the problem of Chapter 10, so, as we have just seen, we have (**??**). Let Q be the program constructed from this P_o in the manner described above, so we have (**??**) and (**??**). Is there any "contradiction" between (**??**) on the one hand and (**??**), (**??**), for this P_o, on the other? There is not. The reason is that (**??**) requires the additional hypothesis X, while we are currently unable to $\vdash X$.

This circumstance suggests that we turn this whole problem around: That we use this "near contradiction" — i.e., (**??**), together with (**??**), (**??**) — to try to establish instead $\vdash X$. But, alas, this doesn't doesnt work either. The reason is more subtle. For (**??**), (**??**), we have that Q is at least as fast as any program that *we can prove* computes P_o, whereas (**??**) provides only faster programs that *actually compute* P_o. How is it that the example, from Chapter 10, only provides faster programs, but not proofs that they are faster? The key lies in the construction of the programs from the P_n, via property 2 above. Each P_n has the same output as P_o except possibly for a finite number of input strings. Consequently, we can obtain, from P_n a program that has *precisely* the same output as P_o. To do this, we consider programs that merely run P_n, but modify its output on a finite number of input strings. One of these, we are guaranteed, will be the P' of (**??**). However, we don't know *which* one. And there is no constructive way to determine which one (for this depends on the halting-behavior of myriad other programs). Thus, not having a specific P' to deal with, we have no way of providing a *proof* that P' has that property. In short, knowing only that there exists an object having a property, there is no direct way of showing that there exists an object that can be proven to have that property. For this reason, then, we are unable to $\vdash \neg X$.

Might it be possible to modify the construction of program Q? Let us say that a program P *effectively computes* the problem of P_o provided: For every input string S, there exists a proof that P, applied to S, yields the same output as P_o, applied to S. We might now wish to modify Q so that it uses, not programs that can be proven to compute the problem P_o, but rather programs that effectively compute the problem P_o. This would be

done by modifying the subroutine $\mathcal{R}$. But, unfortunately, it is not clear how to do this. The difficulty is that there is no way to search for programs that effectively compute P_o, for this notion involves the existence of an infinite number of proofs.

To summarize, we have two things — an example of a problem for which there is no "fastest" program; and a proof that, given any problem there exists a program that is at least as fast as any program that can be proven to compute that problem. These two results, it seems, coexist just fine. This coexistence illustrates a number of points: i) the subtle character of the Blum example, ii) the complications that can arise when one begins to prove things about proofs, iii) how dramatic is the distinction between proving something and having that something "true", and iv) that the strategy of introducing formal sentences, and then $\vdash$-ing them (or failing to), often clarifies an issue.

References

[1] Agrawal, M, Neerja, K, Nitin, S, "Primes is in P", Annals of Mathematics 160, 781 (2004). At http://www.math.princeton.edu/~annals/issues/2004/Sept2004/Agrawal.pdf This paper shows that there exists a program that computes the problem of deciding whether or not an integer n is prime, with difficulty function given by $\leq (\log n)^s$, for every $s > 15/2$.

[2] Arora, S., Barak, B, "Complexity Theory", at http://www.cs.princeton.edu/theory/complexity/ This is a readable summary of (mostly non-quantum) complexity theory.

[3] Blum, M., "A Machine-Independent Theory of the Complexity of Recursive Functions", J. Assoc for Comp Mach 14, 322 (1967). http://portal.acm.org/citation.cfm?coll=GUIDE&dl=GUIDE&id=321395 This paper deals with properties of difficulty that rely only on some general features of the difficulty-measure. There are two very nice results here.

[4] Grover, L, "A Fast Quantum-Mechanical Algorithm for Database Search", in Proc. 28th Annual ACM Symposium on Theory of Computing, ACM, New York (1996).

[5] Hartmanis, J, Hopcroft J.E., "An Overview of the Theory of Computational Complexity", J. Assoc for Comp Mach 18, 444 (1971).

[6] Hennie, F.C., "One-Tape Off-Line Turing Machine Computations", Information and Control 8, 553 (1965). It is proved that no Turing machine can compute the palindrome problem with difficulty function $\ll L(S)^2$.

[7] Karatsuba, A, Ofman, Y, "Multiplication of Many-Digital Numbers by Automatic Computers", Doklady Akad. Nauk SSSR, Vol. 145 (1962), pp. 293294. Translation in Physics-Doklady, 7 (1963), pp. 595596. This paper presents a method to multiply two n-digit numbers with difficulty function $\ll n^2$.

[8] Kelly, J., "General Topology", Springer-Verlag (New York), 1975. An appendix contains the best treatment of axiomatic set theory I have ever seen.

[9] Lenstra, A.K., Lenstra, H.W, eds, in "The Development of the Number-Field Sieve", Lecture Notes in Mathematics 1554, pp 11-42, Springer-Verlag (1993). The sieve method for computing the prime problem is described here.

[10] Mermin, D., "Quantum Computation", Lecture Notes, at http://people.ccmr.cornell.edu/~mermin/qcomp/CS483.html. These lecture notes are a nice, elementary, introduction to this subject. See especially Sect IV.

[11] Pittenger, A., "An Introduction to Quantum Computing Algorithms", Progress in Computer Science and Applied Logic, Vol 19, Birkhauser. This book gives some detail how to do real-world computations with c-not gates.

[12] Shor, P. W., "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer", SIAM J Sci. Statist. Comput. 26, 1484 (1997). This paper proposes a method for using quantum mechanics for efficient factorizing of integers.

[13] Unruh, W.G., "Maintaining Coherence in Quantum Computers", Phys Rev A 51, 992 (1995).

[14] Yasuhara, Ann, "Recursive Function Theory and Logic", Academic Press, 1971. This is my favorite book on Turing machines, unsolvable problems, etc.

Robert Geroch

Unsolvable Problems

1990 Lecture Notes

MINKOWSKI
Institute Press

Robert Geroch
Enrico Fermi Institute
University of Chicago

© Robert Geroch 2013
All rights reserved. Published 2013

ISBN: 978-1-927763-02-5 (softcover)
ISBN: 978-1-927763-05-6 (ebook)

Minkowski Institute Press
Montreal, Quebec, Canada
http://minkowskiinstitute.org/mip/

For information on all Minkowski Institute Press publications
visit our website at http://minkowskiinstitute.org/mip/books/

Contents

1 Introduction 1

2 Games 3

3 Arithmetic 6

4 Encoding 11

5 Procedure 14

6 Will the Technician Stop? 24

7 Appendix 31

1. Introduction

Here is a little game.[1] Begin with string consisting solely of letters "a" and "b",e.g., "ababbbabaa". The leftmost letter of this string is "a". So, append the letters "aa" to the right of the string and delete the leftmost three letters. There results "bbbabaaaa". The leftmost letter of this new string is "b". So, append the letters "bbab" to the right and delete the leftmost three letters. There results "abaaaabbab". Now continue in this way, at each step appending "aa" or "bbab" on the right according as the leftmost letter is "a" or "b", and deleting the leftmost three letters.Thus, in this example the next few steps will yield "aaabbabaa", "bbabaaaa", "baaaabbab", and "aabbabbbab".

What would happen were we continue to play in this way? Clearly, there are just three possibilities. First, we could ultimately arrive at the empty string – the string with no letters – thus terminating the game. This would happen, for example, were the initial string just "aa", for the next step would yield just "a", and the next step the empty string. Second, the string could ultimately repeat itself. Then we would continue to play in a cycle, going through the same sequence of string over and over. This would happen, for example, were the initial string "aabbab", for the next step would yield "babaa", and the next step "aabbab", the same as the initial string. Finally, it is possible that the game go on forever, without terminating or cycling.

More generally, one specifies a tag game by giving the alphabet for the game (in this example, "a" and "b"; but it could consist of more letters), the rules for what is to be added on the right according to which alphabet letter is leftmost (in this example, "aa" if "a" is leftmost and "bbab" if "b" is leftmost; but there could be given any string), and the number of leftmost letters to be deleted (in this example, three; but it could be any number). Then given the tag game and the initial string, one of the outcomes above must occur. How would one determine which outcome will occur in a given case? Well, one could merely play the game. If it terminates or goes into a cycle, one would notice it, and this would be the answer. But what if the game continues, for days, months, decades? Clearly, mere play of the game will not always answer the question of which outcome will occur. Alternatively, one might devise some general argument – not involving actual play – for what the outcome will be. (A simple example is the tag game with alphabet "a" and "b", with "a" $\rightarrow$ "aaaa" and "b" $\rightarrow$ "bbbb", and with three letters deleted at each step. Then every game, no matter what the initial string, goes on without terminating or cycling.)

For any given tag game and initial string one presumably could, given sufficient time and effort, devise some way to determine which of the three outcomes will occur. Let us now ask a more sophisticated question. Can we do more than this? Can we devise some general procedure that will determine, for *any* given tag game and initial string, which outcome will occur?

This last question goes to the heart of the subject of this course. We are interested in well-posed questions (such as, in the example above, the question of whether a given tag game on a given initial string terminates, cycles, or neither), and whether there are procedures for answering those questions (as is asked above). The remarkable result – that has extensive implications both within and without mathematics – is that one can in fact specify well-posed problems that can be solved by no procedure whatever. (In fact, there is no procedure to answer the tag-question above, but the actual example we shell give will be slightly different.)

[1] From Ann Yasuhara's book on recursive functions.

It is our purpose to understand this result. The first step is to gain a somewhat firmer sense of what "well-posed problem" and "procedure" are to mean.

2. Games

Let us consider games of the following type. There are two players – call them "x" and "o". There is a board, visible to both players at all times, the possible configurations of which give the entire status of the game at any one moment. The players alternate turns, and each player in his turn makes a move – makes one of certain allowed changes in the board configuration. Certain board configurations are designated as "x wins", certain as "o wins", and certain as "draw". If, as the game proceeds, one of these configurations is reached, then the game terminates, with the corresponding outcome. If none is ever reached, then the game is declared a draw.

Examples of games that fall within the framework above are tic-tac-toe, checkers, chess, nim (see problem 2 Set 1), and five-in-a-row (Set 1). An example that doesn't is poker (concealed hands violate the board-configuration provision).

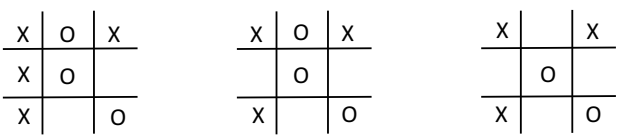
Such games provide good examples of well-posed problems, their solutions, and the issue of whether there are procedures for finding the solutions. The problem is posed to a given player in his turn by the board configuration: He must decide which of the allowed moves to make. The solution to the problem is to make the best move – one that will result in a win if possible, but if not that a draw. More on this in a moment. Whether there is a procedure for solving the problem asks whether one can give explicit instructions such that a technician, merely following those instructions and having no insight, can play expertly, i.e., in every case make the best move (or one of the best, if there are several such).

Let us now detour for a moment to describe in more detail the "best move". Consider the following rules:

1. A board configuration is x-winnable if it is one of the "x-wins" configurations above.

2. A board configuration is x-winnable if it is x's turn, and there is *some* move for x that results in an x-winnable configuration.

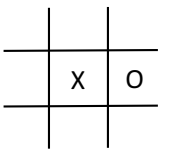
3. A board configuration is x-winnable if it is o's turn, and *every possible* move for o results in an x-winnable configuration. The x-winnable configurations, roughly speaking, are those for which "x, if he plays carefully and skillfully, is assured of winning no matter what o does". In tic-tac-toe, for example, the first configuration below is x-winnable, by rule 1.



Therefore, the second configuration with x's turn, is x-winnable, by rule 2. Finally, the third configuration, with o's turn, is also x-winnable, by rule. (No matter what o does in his turn, there will result a configuration, like that of the second figure above, that is x-winnable.) In a similar way, we define the *o-winnable* configurations. Thus, for example, a configuration is o-winnable if it is an "o wins" configuration; or, if it is o's turn and there is some move for o that results in an "o wins" configuration; or, if it is x's turn and, no matter what move x does, there is some move for o that results in an "o wins" configuration; or, if it is o's turn and there is some move for o such that, no matter what move x then does in his turn, there is some move for o that results in an "o wins" configuration. It continues in this way. In short, an o-winnable configuration is one from which o

can, against any play by x (good play, bad, or whatever), force eventually an "o wins" configuration.

Finally, a configuration that is neither x-winnable nor o-winnable will be called *drawable*. The drawable configurations, roughly speaking, are those from which neither player, with sufficiently skillful play, will lose. In tic-tac-toe, for example, the initial board, before anyone has moved, is drawable. If x takes his first move at the center, then the configuration is still drawable. If, however, o then moves to the side, then the configuration becomes x-winnable. (A move by o to the corner would have made another drawable configuration.)



Suppose that it is x's turn. Let, first, the board configuration be x-winnable before x's move. Then, by rule 2 above, there must exist some move for x such that the configuration is x-winnable after x's move. This would be a good move for x. There may, of course, also be moves for x that result in a drawable or even an o-winnable configuration. These are clearly bad moves. Let, next, the board configuration be drawable before x's move. Then there is no move for x that results in an x-winnable configuration (for by rule 2, were there such a move then the configuration would already by x-winnable, not drawble); and there must exist some move for x that results in a drawable configuration (for were there none then the configuration would already be o-winnable, not drawable). Thus, a good move for x (the best x can do) is one that results in a drawable configuration after x's move. There, may, of course, also be moves for x that result in an o-winnable configuration. These are clearly bad moves. Finally, if the board configuration before x's move is o-winnable, then it doesn't make any difference what x does: The configuration after x's move will in any case again be o-winnable.

We can summarize all this by saying that one can never, on one's move, improve the situation (go from winnable by the opponent to drawable, or from drawable to winnable by oneself). One can always keep it the same. There may also be moves that keeps the situation worse. By a *good* move, we mean one that keeps the situation the same, i.e., that does the best one can do. (You are invited to think through these definitions – of x-winnable, o-winnable, drawable, and a good move – in the case of tic-tac-toe. You should be able to convince yourself that, at least for this game, these terms all take on their usual meanings.)

So, at least for games that fall within our framework, there always exists at one's turn a good move – one that will not cause one's situation (winnable for oneself, drawable, winnable for opponent) to deteriorate. The problem is to find such a good move. For any one specific board situation one could, with some luck and hard work, perhaps come up with a good move. But we now ask for more: We ask for a procedure by which, given the board configuration, one can generate a good move. Is there, given the game, such a procedure?

In the case of tic-tac-toe, there certainly does exist a procedure for finding good moves. Indeed, we all, at least intuitively, know such a procedure. It is not even that hard to write out a procedure in detail. (See problem 1, Set 1.)

Does there exist a procedure for finding good moves in checkers? It is not as obvious as tic-tac-toe, for we are not as skillful checker players. But, I claim, there must exist such a procedure. The reason is that there are just a finite number of possible board configurations in checkers. (This is clear, for each square of a checker board can be in just one of five possible situations – blank, my single piece, my king, your single piece, or your king. Hence, the number of possible configurations of a checker board cannot exceed 5 x 5 x ... x 5 (32 times), which is about 23,300,000,000,000,000,000,000.) Having noticed this, one can, at least in principle, use the following procedure. Purchase a book with 23,300,000,000,000,000,000,000 pages. On each page, draw the corresponding configuration of the checker board (so every configuration is included). Then, at the

bottom of each page, write out a good move for that configuration. This is a procedure in every sense of the word. A technician can well play expert checkers using our book. (At each turn, he merely looks up the board configuration in his book, and makes the corresponding move.) Note that a similar argument could have been used for tic-tac-toe (but was unnecessary there, for we actually knew the strategy).

Thus, we could conclude so far that for every game within our framework there exists an expert strategy – a choice of a good move for every board configuration. (This is essentially by definition of "good move".) Further, in the case of a finite number of board configurations, there always exists a procedure for finding the good moves. A suitable procedure is merely to list every possible board configuration, and for each one, a good move for that configuration. In particular, then, there exists a procedure for expert play in chess.

But what of games with an infinite number of possible board configurations? One such game is nim. (See problem 2, Set 1.) There is a procedure for expert play in this game, as we saw in that problem. In fact, on x's turn, the x-winnable configurations are those for which the remainder on dividing the number of beans in the bowl by three is one or two; and the o-winnable configurations are those for which the remainder is zero. There are no drawable configurations (which figures, since there are no draws in this game).

As a final game, consider five-in-a-row (problem 4, Set 1). This game also has an infinite number of possible board configurations (for the size of the grid on which the marks are made is potentially infinite). Because of this feature, it would be hopeless to try to give a procedure for expert play by writing a "book", because there is no limit to the number of pages that would be required. Does there exist a procedure for expert play in this game? I do not know the answer. The game is very much more complicated than nim, so the rules for expert play would certainly be substantially longer than the rules for nim.

We conclude that, for games within our framework having an infinite number of board configurations, there is no way to assert (as with checkers) that there does exist a procedure for expert play. We may be able to find such a procedure and show it works (as is the case with nim), or we may simply fail to find a procedure, and then not know if there is one at all. So far, at least, we do not have the tools to show in any case that there exists no procedure.

3. Arithmetic

We now turn to a new class of problems – those involving arithmetic. These problems will on the one hand serve to refine our understanding of the notion of a "procedure". Indeed, problems involving arithmetic are better suited than those involving games to illustrate what we wish to mean by a procedure. Further, some of the methods we use here will arise later in our formal discussion of procedures.

By an integer we mean a (positive, negative, or zero) whole number. Thus, 13, –845621, 0 and 2 are integers, while 2.3, 1/2, and π are not. The basic idea is now quite simple: We wish to introduce various questions and problems about integers, and then consider what the solutions are and whether or not there exist procedures for obtaining those solutions.

Is the integer 6 even? We certainly know what the answer is! Is there a procedure to produce the answer? (Recall that a procedure is to be instructions that can be followed mechanically, with no insight, to produce the desired answer.) There certainly is such a procedure. Here is one: "Just say 'yes'. " This is indeed a set of instructions, and it does indeed yield, mechanically, the desired answer. What about the instructions: "Place 6 beans in a pile, and remove them two at a time. If at the end there remains one, say 'no'; if none, then 'yes'."? This is also a perfectly good procedure. But what makes it good is not that it is germane to the problem, but only that it gives the right answer. Thus, the following is also an acceptable procedure for answering our question: "Place 17 beans in a pile, and remove them five at a time. If, when no more can be removed in this way there are still beans in the pile, then say 'yes'." In short, for a procedure we are only interested in whether we get the right answer or not. We are not interested in whether the technician is doing something germane, interesting or whatever. It is only the final answer emitted that counts. The situation here is somewhat analogous to that in tic-tac-toe, in which one can give a procedure for expert play by merely listing every possible board configuration and, for each configuration, a good move. This is an acceptable procedure. It is not necessary, for a good procedure, that the technical actually do anything germane to tic-tac-toe (such as make arguments about "if I do this, then my opponent will do this..."). In short, returning to our original question, a suitable procedure for answering it is any explicit instructions for a technician that result in his producing eventually the answer "yes".

Is the number 7 even? All the remarks above apply, and we have only to produce instructions that result in "no".

Given integer between 1 and 10, is it even? It will be recognized that this is not a single question, but rather is ten questions ("Is 1 even?", "Is 2 even?",..., "Is 10 even?") all assembled together. So, there is not a single answer, but ten answers, one of each choice of integer. There clearly is a procedure for answering any one of these questions (and, indeed, we have just done two of them, above). But now we ask for more. We ask for a single procedure, given once and for all to our technician, such that our technician can, armed with that procedure, answer all ten questions correctly. Let's try a few possibilities. Consider "Just say 'yes'." This is certainly a procedure, but it is not a correct procedure for our problem (for, e.g., it gives the wrong answer if "7" is presented). Our procedures, then, had better require that the technician actually use the number presented. Here is a suitable procedure: "If the given integer is 2, 4, 6, 8, or 10, then say 'yes'; if 1, 3, 5, 7, or 9, then say 'no'." Note that this does in fact produce the correct answer – and that is all we care about in the procedure. Another suitable procedure, for

example, is "Place in a bowl a number of beans equal to the given integer, then remove them two at a time until this is no longer possible. If a bean then remains, say 'no'; if none remain, say 'yes'." But, again, the procedure need not be germane to the problem. Thus, here is a suitable procedure: "If the number, when spelled out, either ends in an 'e', or contains two 'e"s, then say 'no'; otherwise, say 'yes'." Again, the situation here is similar to that of tic-tac-toe: Any procedure that results in expert play is ok, whether or not that procedure gives any particular insight into the game itself. In particular a mere list of board configurations and corresponding good moves is acceptable.

Given any positive integer, is it even? This will be recognized as not one question, not ten questions, but rather an infinite number of questions. Our previous methods for obtaining procedures do not work in this case. "Just say 'yes'." does not, of course, always produce the correct answer. And a listing of all possible integers that might be presented, together with the answer to give for that integer, is not possible with an infinitive number of possible integers. Is there, then, a procedure for answering these questions? There certainly is. One procedure is bean-counting: "Place in a bowl a number of beans equal to the integer given, and remove them two at a time until that is no longer possible. If a bean then remains, say 'no'; if not, say 'yes'." In fact, there are easier procedures. For example: "If the given integer ends in 2, 4, 6, 8, or 0, then say 'yes'; if in 1, 3, 5, 7, or 9, then say 'no'." Note that this last procedure "just happens" to work. Being even is quite different from ending in 2, 4, 6, 8, or 0. But it just happens to be the case that the two are equivalent. This example just further illustrates our point: All that matters of a procedure is that it produce the right answer, not whether the details of that procedure are relevant to the problem, or interesting, or anything else. The situation for this question is rather like that of the game of nim. In the game, there are an infinite number of board configurations, in analogy with an infinite number of questions here. Yet, despite the infinite numbers in both cases, we can find a procedure – for expert play in nim, and for answering the questions here.

Given a positive integer, is it evenly divisible by 3? This question is rather like that above, except that we need for it a somewhat a different procedure. Divisibility by 3 is not revealed by the final digit alone. We could, for example, use bean-counting. Or, one could write out a procedure that includes an explanation of how to do long division (for that, after all, is itself a procedure). Then, our procedure would dictate that the given number be divided by 3, and the remainder examined. Of course, our instructions for long division would have to include the multiplication table, but that is ok: We merely include it.

For the next few examples, we need the following definition. An integer two or greater is called *prime* if it is evenly divisible only by the integers one and itself. Thus, for example, 2, 3, 7, 19, 97, and 5737 are prime integers, while 4, 6, 35, 111, and 5767 are not.

Given a positive integer, is it prime? This is, of course, an infinite number of questions. There is, as it turns out, a procedure for answering them. Suitable instructions might direct that the technician divide the given integer in turn by each integer from 2 up to one less than the given integer, and find in each case the remainder if any. If in every case there is a remainder, then answer "yes"; if in any case there is none, then answer "no". Thus, for example, were the given integer 13, then one would try to divide it, in turn, by 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12. In each case one would obtain a remainder, and so one would answer "yes". Of course, these instructions for our technician would also have to include detailed instructions for how to perform long division (including, e.g., the multiplication table), but there is no difficulty in including this. Alternatively (or actually, if you think about it, the same thing), we could direct that the technician write down

in order the integers from 1 up to the given integer. Cross out every second in this list, then every third, every fourth, and so on up to one less than the given integer. If the given integer is not in this process crossed off, then say "yes"; if it is, say "no". So, for example, if the given integer were 13, the technician would write 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. He would then cross off every second number (namely 2, 4, 6, 8, 10, 12), then every third (namely 3, 6, 9, 12), and so on, up to every twelfth. In this process, "13" would never be crossed off, and so the technician would answer "yes".

Given a positive integer, is it the sum of any two prime integers? There is a procedure for answering this question. Suitable instructions might direct that the technician consider in turn each integer from one up to one less than the given integer, and determine whether that integer, and the given integer minus that one, are both prime. If in any case they are both prime, answer "yes": if never both prime, answer "no". Of course, the technician could determine whether these various integers are prime, using the procedure we just described above. Thus, for example, if the original integer were 16, then the technician would ask "Are 1 and 15 both prime?", "Are 2 and 14 both prime?", 'Are 3 and 13 both prime?", and so on. In this example, one would find a case in which both are prime (namely, 3 and 13), and so the technician would answer "yes". On the other hand, were the given integer 11, the technician would, as one can check directly, ultimately answer "no".

Given a positive integer, is it the difference of any two prime integers? This question is of course quite similar to the one above, and so one might try a similar procedure. Thus, one might direct that the technician consider in turn each integer beginning with one more than the given integer, on up, and determine whether that integer, and that integer minus the given integer, are both prime. If in any case they are both prime, then answer "yes"; if never both prime, answer "no". Thus, for example, if the original integer were 16, then the technician would ask "Are 17 and 1 both prime?", "Are 18 and 2 both prime?", "Are 19 and 3 both prime?", and so on. In this example, one would find a case in which both are prime (namely 19 and 3), and so the technician would answer "yes". Yet this is not a procedure! The problem is that the technician is being asked to test an infinitive number of cases (from one more than the given integer "on up"). It all works very well if, as in the example above, a case is found with both prime, for then the technician can stop and say "yes". But what if no such case is found – after a day, after a month, after decades? When will the technician be directed to say "no"? Indeed, if you have a lot of time on your hands, you may wish to play technician, say, starting with the given number 19. In short, without a guarantee that the technician's efforts will at some point come to an end, this is not a procedure as we wish to use this term. The key difference between this example and the preceding one, of course, is that before there were only a finite number of cases to be tested for any given integer, whereas here there could be an infinite number. So, our given candidate for a procedure fails. That does not mean that there exists no procedure, only that we have not found one yet. I do not know whether or not there exists such a procedure.

Given a positive integer, is it the largest prime? This question appears on its face to be very similar to the previous question. Thus, the obvious "procedure" here would be to test for primeness each positive integer starting with one more than the given integer, on up. If and when one is found to be prime, answer "no". But this, as it stands, may not be a procedure, for we have as yet no guarantee that the technician's task will not go on forever, But what sets this question apart from the previous one is the following:

Theorem. There is no largest prime integer.

Proof: Let n be any positive integer. We show that there is a prime integer

larger than n. Consider the number z = 1 x 2 x 3 x ... x n + 1. That is, z is the result of multiplying 1 times 2 times 3, and so on all the way up to the given integer n, and then adding 1 to the result. Now z is either a prime integer or it is not. If it is prime, then – since z is certainly larger than n – we have indeed found our prime integer larger than n. But what if z is not prime? Then by definition, we may write z as a product of two integers. If either of these is not prime, we may in turn write it is a product of two other integers. Substituting, we will have an expression for z as a product of three integers. Continuing in this way, always replacing any non-primes in our expression by products of other integers, we will eventually end up with an expression for z as a product of prime integers. What prime integers can appear in this final expression? Well, 2 cannot appear, for z is not evenly divisible by 2. (Indeed, z was originally given as a product of 2 and something (namely 3 x 4 x ... x n) plus 1. So, the reminder on dividing z by 2 is 1.) Similarly, 3 cannot appear in this final expression; nor can 4; and so on up to n. That is, the prime integers that appear in our final expression must all be greater than n. Thus, we again find a prime integer greater than n.

What we have shown, then, is that given any positive integer n – no matter how large – there exists a prime integer larger than n. Clearly, then, there can be no largest prime integer.

Armed with this result, let us now return to our original question. We can claim, now, that what we originally gave for this question was indeed a procedure. Our theorem guarantees that the efforts of the technician will indeed terminate, and so he will indeed eventually produce an answer (in fact, "no"). But there are, of course, alternative procedures for this question. We could, for example, tell the technician "Just say 'no'." What we learn from this example, then, is that a set of instructions, to be a procedure, need only have the property that it does indeed terminate and lead to the correct answer. It need not be obvious to the technician that the instructions will terminate.

Given positive integer n, do there exist positive integers x, y, and z, such that $x^n + y^n = z^n$? (That is x raised to the n^{th} power, plus y raised to the n^{th} power, must equal z raised to the n^{th} power.) If, for example, n = 2, then the answer is "yes". Indeed, set x = 3, y =4, and z = 5. We have $3^2 = 9$, $4^2 = 16$, and $5^2 = 25$, while 9 + 16 = 25. For n = 3, the answer, it turns out, is "no" (but the proof is somewhat difficult). For general n, the answer is not known. Is there a procedure to answer this question. The obvious method would be, given n, to try various x's, y's, and z's. But this does not look promising, for we have no way to guarantee that the process terminates. Thus, it seems difficult offhand to come up with any suitable procedure. But note that, on the other hand, we are not in a position to assert that no procedure exists. There is a fair chance, for example, that the only n's that work are n = 1 and n = 2, all others failing. Then, suitable instructions would be "If n is either 1 or 2, say 'yes'; otherwise, say 'no'." Here, then, is an example of a question such that we just do not know whether or not there exists a procedure for answering it. Some day, perhaps our knowledge of mathematics will have risen to the point that we do know. We emphasize, however, that, for a given question, there either is a procedure or there is not. It is not a function of our state of knowledge in mathematics. As we learn more mathematics, we will merely know in more cases whether or not there exists a procedure.

It is true that, for all n greater than 2, there are no positive integers, x, y, and z, such that $x^n + y^n = z^n$? It is not presently known that the answer to this question is. Does there exist a procedure for answering it? Note that this is a single question: The answer is either "yes" or "no". Thus, there certainly does exist such a procedure. It is either the one with instructions "Say 'yes'.", or the one with instructions "Say 'no'." Of course, we do not happen to know, currently, which of these is the correct procedure. Perhaps we will some day. But in any case

we do know that one of them is, and so we know that there *exists* a procedure to answer this question. In short, we wish to distinguish between the existence of a procedure for answering a question and the act of actually answering the question. We wish to allow the existence of a procedure if we can show it exists, whether or not we happen to have it in hand.

Given positive integer n, what is the n^{th} digit in the decimal expansion of π ($= 3.1415926535\ldots$)? (Thus, for example where n given as 7, then the correct answer is "6", the seventh digit in the above expansion.) There does in fact exist a procedure for answering this question. There are available any number of formulas that allow one to compute the value of π to any desired degree of accuracy.

Given integer from 0 to 9, does it occur ten times in a row in the decimal expansion of π? We do not happen to know the answer to this question for any value of the integer. Nonetheless, there is a procedure to answer the question. We can basically consider all possibilities. Thus, we have "Say 'yes'.", "Say 'yes' for 0, 3, 5, and 6; and 'no' for 1, 2, 4, 7, 8, and 9.", and so on. There is a grand total of 1024 such possibilities. Since each of these ten questions has an answer (either 'yes' or 'no'), one of these possibilities is the correct procedure. Hence, there *exists* a procedure to answer the question.

Given any positive integer, does it recur ten times in a row in the decimal expansion of π? (Thus, for 37, we ask whether "37373737373737373737" appears anywhere in the expansion.) The above procedure will not work here, for there are an infinite number of questions, and so we cannot make a list of answers. I do not know whether there is a procedure for answering this question. It may, for example, turn out that the answer is always 'no', or always 'yes', in which case there certainly does exist a procedure. On the other hand, it is at least conceivable that the yes's and no's are so scattered around in the answers that no procedure will reproduce them.

4. Encoding

There is clearly an enormous variety of questions with respect to which one could ask for procedures. A first step in understanding procedures better would be to standardize these questions. (Ultimately, indeed, we shall even standardize the procedures themselves.) For example, some of our earlier questions required as input a board configuration; those more recent, a positive integer. A question such as "Given positive integers a, b, and c, does a + b = c?" requires as input several integers. It would be helpful, then, if we could have a standard input – as opposed to all this variety – for our questions. This standardization is carried out by encoding various types of information into integers.

We seek first a method for encoding two positive integers into a third. That is, we wish a rule that allows us, given the two integers, to compute a third – but this rule must be such that, given only the third integer, one can "decode" to determine the original two. (Thus, for example, adding the two integers would not do, for one could not in general recover the two addends given only their sum.) We shall adopt the following rule: Write the two integers one after the other, but between them insert a number of "0" 's equal to the total number of digits of the two integers combined. Thus, for example, were the integers 30060 and 72, then we would encode these to obtain the single integer 30060000000072. (Note that we insert 7 "0" 's between the two integers, since those integers together have 7 (= 5 + 2) digits.) This rule indeed satisfies our criterion. To decode, take the total number of digits of the given number, divide by two, find the first that many "0" 's in a row, starting from the right, and cross them out. There results the original two integers, Thus, for example, given 1900000300, there are ten digits. Removing the first five o's in a row, starting from the right, we are left with '19 300". These are the original two integers. Of course, there are many integers that cannot be obtained as a result of encoding others, e.g., 2064 and 2000064.

Given positive integer n, is it the result of encoding positive integers a and b, with a greater than b? There is, of course, a procedure to answer this question. But note the role that encoding plays. This is really a question about two integers (a and b), but, by use of encoding, it has as its input a single positive integer, n.

How would one encode three positive integers into a single integer? There is a simple way. First encode the second and third, to obtain a new positive integer. Then encode the first and that integer. Thus, for example, let the three numbers be 111, 60, and 21. We first encode the second and third, to obtain 60000021. We then encode 111 and 60000021, to obtain 1110000000000060000021. From this single integer we can, in turn, recover our original three. Indeed, we first decode 1110000000000060000021, to obtain 111 and 60000021. Then we decode the latter, to obtain 60 and 21. Thus, we recover our three integers, 111, 60, and 21. Note that there will never be any ambiguity in this, as long as one knows the code.

Similarly, to encode four positive integers, we first encode the third and fourth, then the second with the result, then the first with the result. There results a positive integer that can be decoded to obtain the original four. It works similarly with any given number of positive integers.

Given positive integer n, is it the result of encoding three positive integers, a, b, and c, with a + b = c? There is of course a procedure. By encoding, this question (actually about three integers) is made to require as input just a single integer.

In each of the examples above, we are given, beforehand,

the number of integers that are being encoded. Thus, for 1110000000000060000021, if we knew it was to be decoded into three integers we would obtain 111, 60, and 21; while if we knew it was into two integers we would obtain 111 and 60000021. Is there some way we can arrange that not only the integers themselves, but also the number of integers, is included in the encoding process? That is, we wish to be given a single integer – and no other information – and obtain from it both how many other integers are to be obtained, and what they are. This is actually quite easy. Let there be given any list of positive integers we wish to encode. Count the number of integers in this list, append this number to the beginning of the list (to obtain a new list, with one more integer), and finally encode this new list as above. Thus, for example, let the original list of integers to be encoded consist of 13, 9, and 6. There are three integers in this list, and so our new list consists of 3, 13, 9, and 6. Finally, the result of encoding these four integers is 3000000000000130000009006. It is clear that all this can be decoded. Thus, in the example above, given 3000000000000130000009006, we first decode once, to obtain 3 and 130000009006. The first integer, "3", tells us how many integers remain encoded in the second integer, 130000009006. Thus, we decode this twice more, to obtain finally '13", '9", and "6" our original three integers. Try with 200000000010300005.

Given positive integer n, is it the encoding of a list of two or more integers such that the last is the sum of the others? Here is a question about an indeterminate number of integers that, through encoding, is rendered as a question about a single integer.

Much more detailed information could also be encoded into a single positive integer. Consider, for example, the novel *War and Peace*. One could first assign numbers to the various symbols that appear in the book, e.g., "a" = 1, "b" = 2,..., "z" =26, "." = 27, "?" = 28, " " = 29, and so on. Since the book itself is merely a sequence of such symbols, we could replace it, first, by the corresponding sequence of integers. Thus, if the book begins "It was a ...", our sequence would begin 9, 20, 29, 23, 1, 19, 29, 1, ... Now, we simply take the resulting sequence of integers, and encode it as above (using the final version, for an indeterminate number of integers, since we do not know, a priori, how many there are to be) into a single integer. *War and Peace*, in this way, is rendered as a single positive integer.

One might also encode information about the various games discussed earlier. Consider, for example, tic-tac-toe. Let us number the squares of the board as shown on the right. Then we may represent a given board configuration by a nine-digit number, where each digit represents the corresponding square. Each digit, in turn, will tell us what is in that square, say using the following code: "x" = 1, " " = 2, "o" = 3. Thus, for example, the board configuration shown on the right would be encoded as the integer 122312223. Clearly, one can decode to determine the board configuration from the integer. One could also include whose move it is, e.g., by including a tenth digit on the right that is "1" if it is "x" 's move, and "3" if "o" 's. Thus, the above configuration, with "x" 's turn, would be represented 1223122231.

1	2	3
4	5	6
7	8	9

Given positive n, is it the encoding, above, of a tic-tac-toe board configuration and whose move it is such that this configuration is x-winnable? Drawable? These questions about tic-tac-toe thus, by encoding, become questions about positive integers.

X		
O	X	
		O

We could represent a "move" in tic-tac-toe by giving a 19-digit number consisting of the initial board configuration (first 9 digits), whose turn it is (next digit) and the final board configuration (last 9 digits). Thus, if it is x's turn above and x takes the upper right hand corner, this

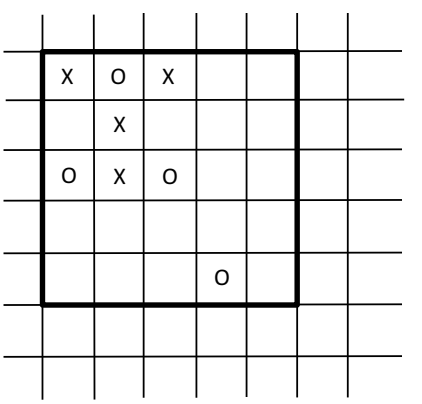
move would be represented as 1223122231121312223. Clearly, one can decode to recover the move in the more familiar form.

Given positive integer n, is it the encoding, above, of a good move in tic-tac-toe?

In a similar way, one could encode positions and moves in checkers. (Number the squares on the board, assign numbers to represent what can be on those squares, and then represent the board configuration by some large integer, etc.) One could encode a nim move by first determining the number of beans in the bowl before the move, whose turn it is (say, "x" = 1, "o" = 3), and the number of beans in the bowl after the move. These three integers could then be combined, using our encoding above, into one. Thus, for example, 28000000300027 is a good move in nim, while 28000000300026 is not.

Finally, let us consider five-in-a-row. The complication here is that, as compared with checkers, say, the board is infinite. We may represent a board configuration in five-in-a-row as an integer, for example, as follows. Given a board configuration, first draw on the board a large square that includes all moves made so far. Then number the small squares within that large square from left to right, top to bottom, as in tic-tac-toe. To encode the board configuration, first give the integer that represents the length of a side of the square, and then the integers that represent what ("x" = 1, " " = 2, "o" = 3) is on each small square. Then take all these integers, and encode them into a single integer as above.

Consider, as an example, the five-in-a-row board configuration at the right. All the moves can be encompassed in the 5 x 5 square shown. Thus, in the first step we represent this board configuration by a total of 26 integers, namely 5, 1, 3, 1, 2, 2, 2, 1, 2, 2, 2, 3, 1, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2. These 26 integers would then be encoded, using our earlier procedure, into a single positive integer. Similarly, then, moves, etc. in five-in-a-row could be encoded into integers. The question of whether a five-in-a-row configuration is o-winnable, or whether a move is a good move, can thus become a question about positive integers.



What is important in the above is not so much the details of the various encoding procedures. Rather, it is the following idea: Any finite body of information can be suitably encoded into a positive integer.

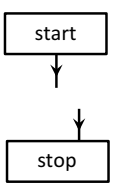
5. Procedure

We now have available a standard input for our questions: We may take as the input an integer, secure in the conviction that we can always encode what we really want to ask about into such an integer. It is time now to standardize the procedures themselves.

The idea is to supply the technician with two things: a scratch pad, in which the technician can record various intermediate results for later use as well as the final answer; and a set of instructions, which will direct the technician in explicit detail what to do.

For the scratch pad, we shall use a number of bowls, each capable of holding an arbitrary number of beans (9, 1, 2, ...). At any moment, a bowl must hold some specific number of beans (say, "13" – It cannot hold "an infinite number of beans".) We may indicate the bowls by drawing them on paper, or, more easily, by giving each one a name, e.g., "Nancy",, "C", "137", or "* ". It is only necessary that we (actually, the technician) be able to identify and distinguish these names. Any given procedure may use but a finite number of bowls (so the technician can be guaranteed that, by hunting around, he can always find the appropriate bowl).

For the instructions, we shall proceed as follows. We first introduce seven basic instructions. We then string these together, to make more complicated sequences of instructions, by use of flow charts.

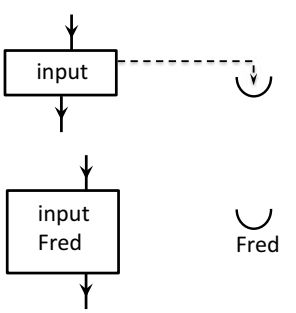


The seven basic instructions are the following:

The instruction "start" has only a flow line emerging from it. The technician will always begin by finding, and then starting from, a "start" instruction.

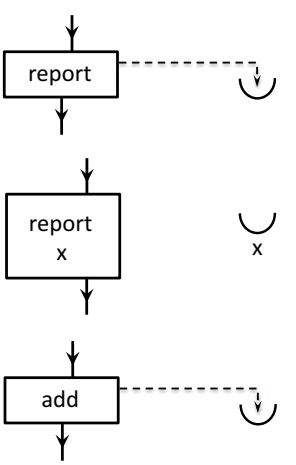
The instruction "stop" has only a flow line entering into it. The technician, on reaching the instruction "stop", stops.

The instruction "input" has a single flow line entering into it, and single line emerging from it. In addition, this instruction must indicate one of the bowls comprising the scratch pad. This could be done, for example, by drawing a little dashed arrow to the bowl, or by giving the name of the bowl. On reaching the "input" instruction, the technician waits for us (who are overseeing all this) to place within the indicated bowl some number of beans of our choosing. Once this has been done, the technician continues. The role of this instruction is to allow us to supply information to the technician.



The instruction "report" has a single flow line entering into it, a single flow line emerging from it, and also indicates one of the bowls. On reaching the "report" instruction, the technician announces to us the number of beans then in the indicated bowl, and then continues. The role of this instruction is to allow the technician to supply information to us.

The instruction "add" has a single flow line entering into it, a single flow line entering into it, a single flow line emerging from it, and also indicates one of the bowls. On reaching the "add" instruction, the technician drops one additional bean into the indicated bowl, and then continues. This instruction, which will



ultimately allow the technician to "calculate", is automatic: We play no role in it.

The instruction "remove" has a single flow entering into it, and also indicates one of the bowls. On reaching the "remove" instruction, the technician takes one bean from the indicated bowl (That is, if there is a bean in that bowl. If the bowl was empty, the technician leaves it empty.), and then continues.

The instruction "empty?" has a single flow line entering into it, and *two* flow lines emerging from it (these marked "yes" and "no"), and also indicates one of the bowls. On reaching the "empty?" instruction, the technician peers into the indicated bowl to see whether or not it contains any beans. If the bowl is empty, the technician continues along the flow line marked "yes", if not, along the line marked "no". The number of beans in the bowl is not thereby changed.

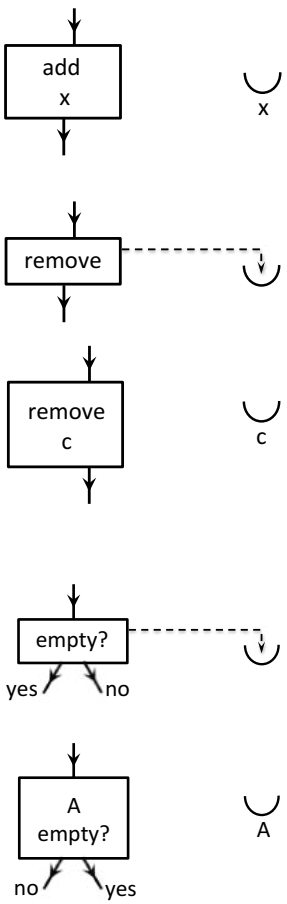
These, then, are the seven instruction: two to begin and end, two to communicate with us, two to calculate, and one to "make decisions".

We next use these seven instructions as building blocks, composing more complicated tasks from them by means of flow charts. The rules for flow charts are the following:

1. Every flow chart must have one and only one "start" instruction (for otherwise the technician wouldn't know where to begin).

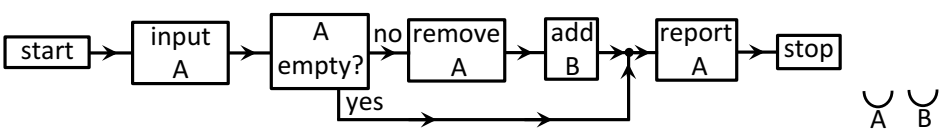
2. When several flow lines meet at a node, , any number of lines may come in, but one and only one flow line may emerge (for otherwise the technician, on reaching the node, would not know which way to go).

3. A flow line cannot just come to an end (for otherwise what is a technician to do?).



Let us also agree that initially, when the technician starts, all the bowls are to be empty.

An example of how these instructions operate is provided by the following flow chart:

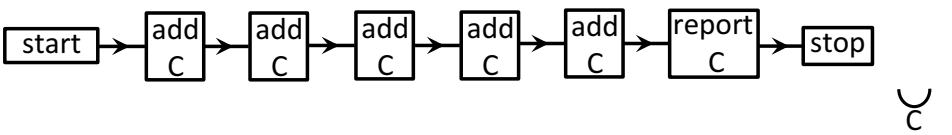


First note that each instruction has the correct flow lines entering and/or emerging from it, as well as appropriate bowl indications, and that the flow chart itself satisfies our rules. Here, the scratch pad has two bowls, named "A" and "B". A technician, following this flow chart, would proceed as follows. Beginning at "start", he would first ask us to place some number of beans in bowl "A". He would then test to see if this bowl is empty (i.e., if we put in zero beans). If not, he would remove a bean from bowl "A", and place it in bowl "B". Then, in either case, he would report to us the number of beans in bowl "A", and finally stop.

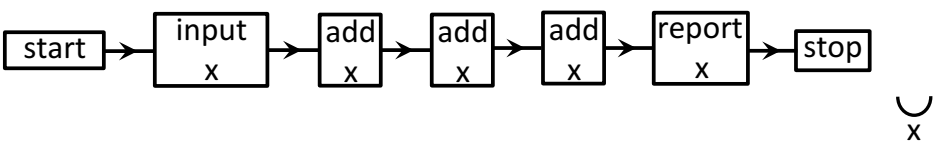
We now intend to argue in favor of the following assertion: *Such flow charts, with such instructions and such scratch pads, represent precisely our intuitive understanding of the word "procedure". That is, any task represented by such*

a flow chart would be regarded as a procedure; and any task we would regard as a procedure can be represented by a flow chart. You probably would agree already with the first assertion (that any task represented by a flow chart would be regarded as a procedure), but are more skeptical of the converse. In any case, our argument will consist of giving various examples of flow-chart tasks. (It is almost impossible, by the way, to learn this subject passively. You must draw your own flow charts, invent your own, and play technician on your own.)

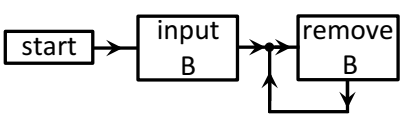
The flow chart below places five beans in bowl "C", and then reports the result, "5":

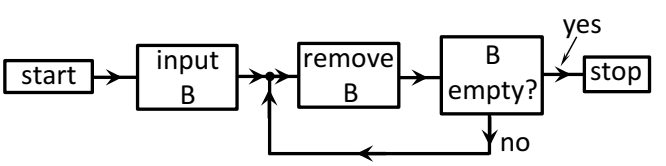


This flow chart inputs a number, adds three to it, and then reports the result:



The first flow chart below empties the bowl B. However, this way of accomplishing the task is an inconvenient one, for it ties up a technician indefinitely. (Even after B is empty, the technician continues, forever, to try to remove additional beans.) A more convenient version is the second. We ask the technician to test B for emptiness after each bean-removal.





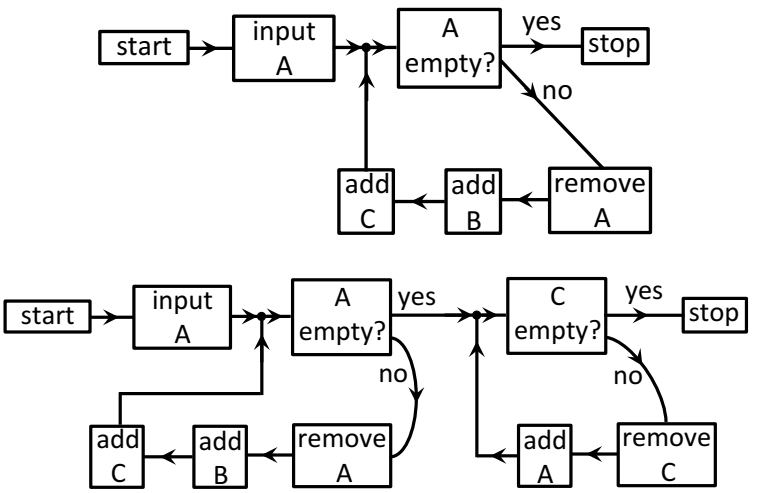
The flow chart below "copies" the contents of bowl A into bowls B and C, at the same time depleting A to empty.

Thus, for example, if at the "input A" we were to place in A seven beans, then we would wind up, at "stop" with seven beans in each of B and C, and zero beans in A. How would one make a copy of A without emptying it? It is simple: One first copies A into B and C, while emptying A, as above; and then one copies C back into A:

Similarly, we can make any number of copies of any bowl. This is most useful, for it allows us to employ our scratch pad to "keep records".

The flow chart below adds A and B, storing the sum in bowl A and emptying the bowl B:

Suppose, however, that we wished to place the sum of the numbers in A and B in a new bowl, C say, while restoring A and B to their original numbers? This could be accomplished by combining the "add" and "copy" flow charts above: First copy A into C and B into D. Then add the numbers in C and D, with the sum appearing in C, as above.



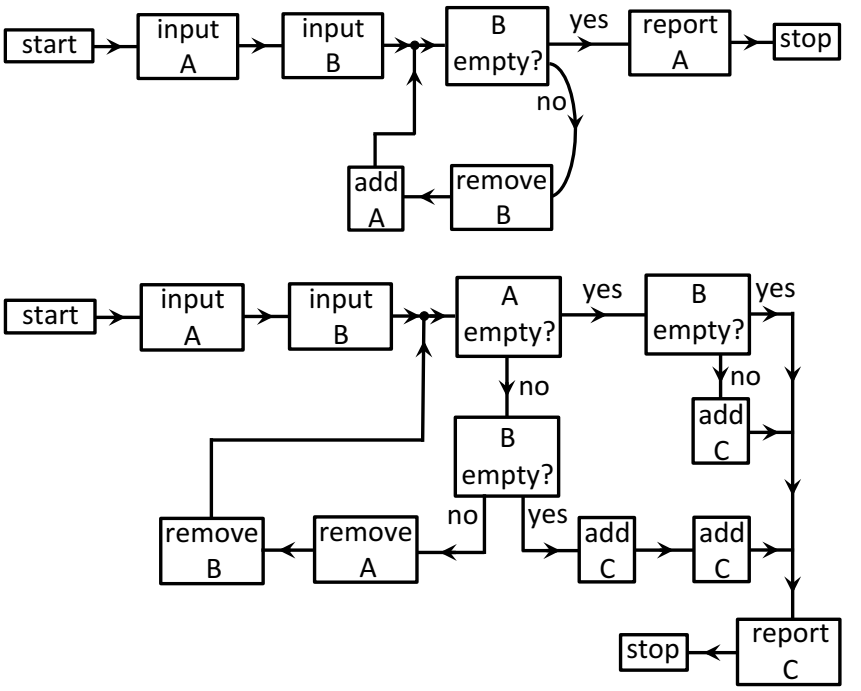
The flow chart below takes two nonnegative integers, A and B, and determines whether A is greater than B, equal to B, or less than B:

Here, the answer is encoded in the number of beans in bowl C, which is reported just before the end. The code is: 0 beans for A and B equal, 1 bean for A less than B, and 2 beans for A greater than B. Combining this with the "copy" flow chart, we could further arrange that the bowls A and B be restored to their original numbers at the end. Similar to this one is the flow chart that subtracts B from A (reporting zero if B is greater than A):

Suppose that one had wanted something something different reported when B is greater than A, say B minus A? This could be achieved by combining this flowchart with the one above: One would first employ the earlier flow chart to determine whether A is greater than, equal to, or lesser than B, and then require that different instructions be carried out depending on the outcome.

Multiplication is "repeated addition". Based on this idea, we construct the following flow chart to multiply nonnegative integers A and B:

The dashed box in this flow chart is not one of our seven basic instructions. Rather, it stands for that combination of those instructions that carries out the following task: Cause C to contain at the end a number of beans equal to the sum of the numbers originally in A and C, and leave the number of beans in A unchanged. This is something we already know how to accomplish by means of our seven basic instructions. In more detail, then, the dashed box stands for the following:



Such abbreviations are a great convenience, of wish we shall make extensive use. In using such abbreviations, it is important that one i) indicate clearly within the dashed box what is to be done, and ii) be capable, if pressed, of actually doing this using the seven basic instructions. Let us also agree on the following: by "place... in bowl C" we mean "cause C at the end to contain exactly... beans"; and all bowls, unless otherwise indicated, are to be restored, within the dashed box, to their original numbers of beans. As a second example of such abbreviations, consider the following flow charts:

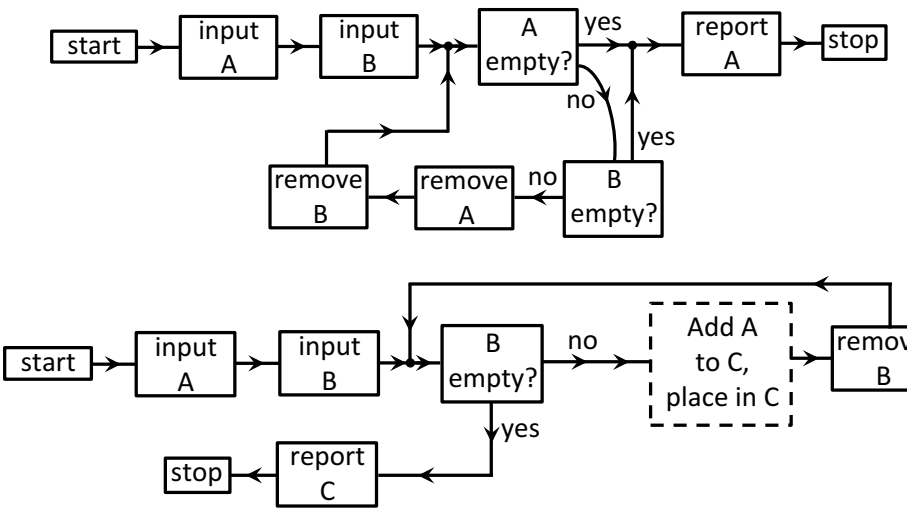
Here, the dashed box places in bowl C the product of the numbers of beans in A and C, restoring A to its original number. This is something we already know, from the discussion above, how to do. What does this flow diagram do? It raises A to the power B. (It is practically the same as the "multiplication" chart, but with multiplication now replacing addition.) Note that we must begin with one bean in bowl C.

As a final example, we give a flow chart for division. Here, A is divided by B, with there reported first the quotient (in C), and then the remainder (in D).

Here, for the first dashed box there is to be determined whether or not A is less than B, and, with bowls A and B restored to their original values, the appropriate exit path taken. One could, by means of our basic instructions, carry this out. (One would use, basically, a modification of the second flow chart on page 17.) what will happen if, for "input B" there were placed zero beans in bowl B? How would one modify the flow chart to make something different happen?

We have now displayed flow charts to carry out certain bookkeeping operations on integers – copying and comparing – as well as the basic operations of arithmetic – adding, subtracting, multiplying, and dividing. More complicated operations are now constructed from these. Some examples follow.

The flow chart above determines, provided A is at least two, whether or not A is a prime integer. This answer is encoded in the final report (which can, of course, only be an integer) as follows: 0 if prime, 1 if not prime. The flow chart operates, basically, by testing, for each number B from 2 up to one less than A, to see if it divides A without a remainder. What happens if A is zero or one?



The next flow chart, given above, answers the questions of whether or not A is the sum of two primes.

The answer is reported as an integer, with "0" for yes and "1" for no. Alternatively, one could ask whether A can be written as the difference of two primes:

But notice an important difference between these two flow charts. In the first chart, the efforts of the technician cease in any case, with the report as to whether or not A is the sum of two primes. In the second chart, by contrast, the technician stops and reports only in the case in which A is found to be the difference of two primes. If not so found, the technician's efforts continue forever. This difference reflects, of course, a difference between the two questions. There exists a procedure to answer the first, but not, as far as we know, the second.

The next flow chart finds, for B a positive integer, the B^{th} digit (counting from the right) of A:

What does this flow chart report if B exceeds the number of digits in A?

Similarly, we may have reported the total number of digits in A:

Once we can "count digits", we can write the flow chart that takes positive integers A and B, encoding them, according to our prescription, into a single integer:

The next flow chart decodes:

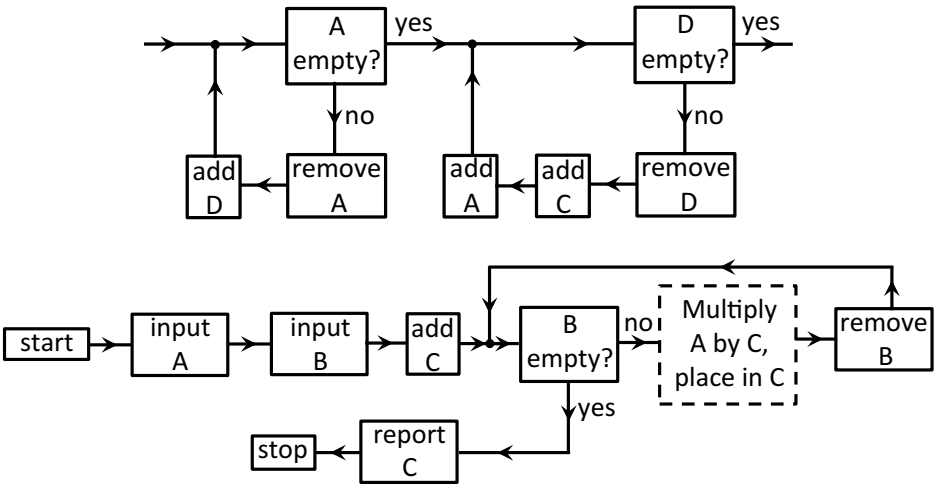
In case A is not the encoding of two positive integers, the flow chart stops, without reporting anything. If A is such an encoding, those two numbers (in bowls F and E) are reported before stopping. The encodings of other types of information are treated similarly. For example, the flow chart below determines whether the integer A is the encoding of some tic-tac-toe board configuration:

The number "1" is reported for yes, "0" for no.

The flow chart below determines whether or not there appears a "4" among the first A digits in the decimal expansion of π:

The answer, again, is reported as "1" for yes and "0" for no. Here, the dashed box marked "place in C the B^{th} digit of π" represents a certain (very complicated) flow chart that uses, for example, one of the series expansions for the number π. The details of this chart are not of interest here. Alternatively, one could ask whether "4" appears after the A^{th} digit:

But notice, in contrast to the previous one, that this flow chart can only report "yes" (i.e., "1"). If no digit "4" is found, then the technician will continue forever within the flow chart. Again, this behavior reflects the fact that we as yet have no procedure to answer this question, whereas we do have a procedure for the prior

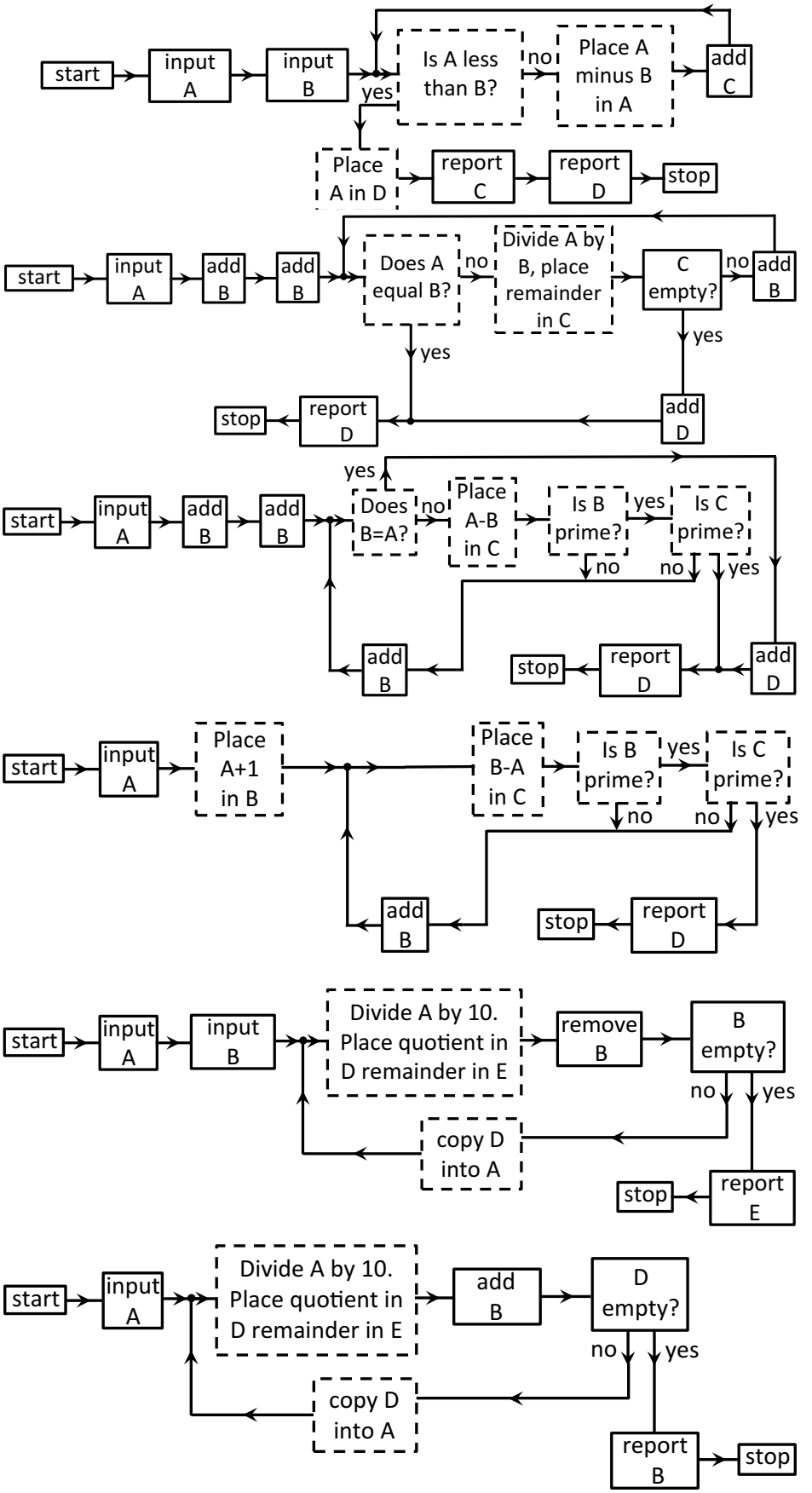


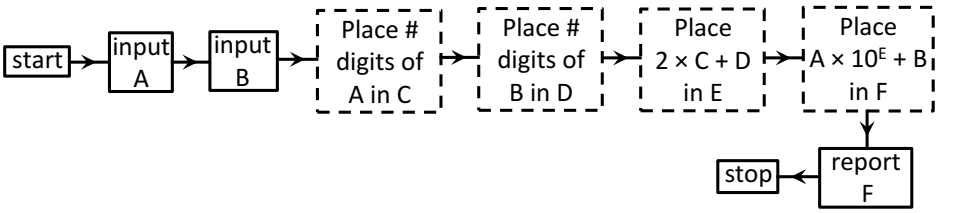
one.

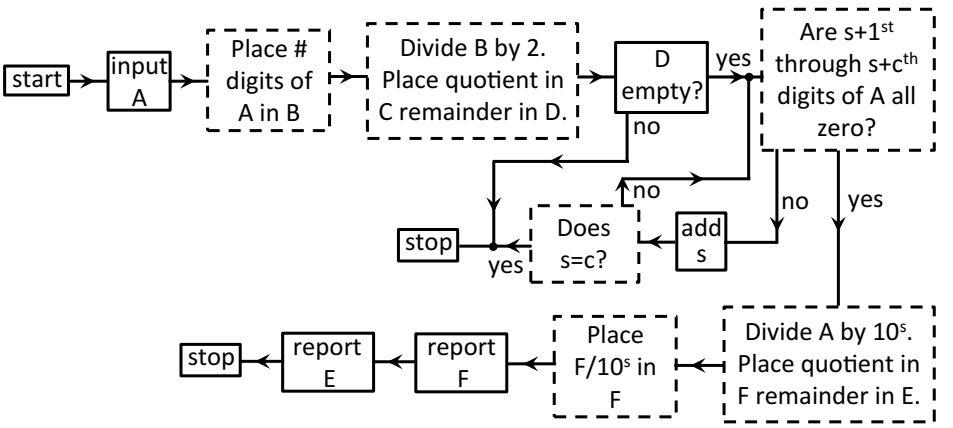
As a final example, we give the flow chart that asks whether, given positive integer n, there exist positive integers x, y, and z with $x^n+y^n = z^n$. Here, we must "try various choices of x, y, and z", and one way of doing this is to "try choices of a single integer A, asking whether it is the encoding of three such integers".

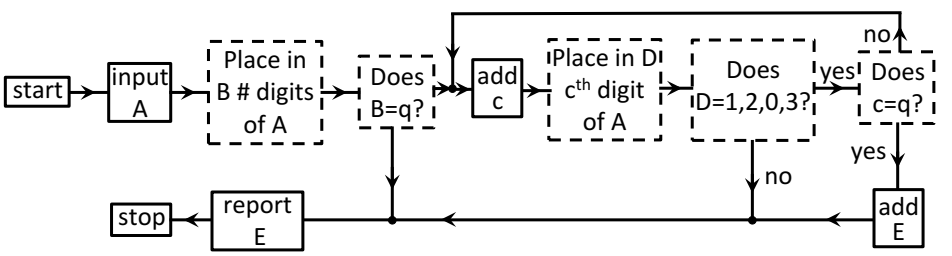
Here, the technician reports "0" and stops if such x, y, and z, are found, while the technician continues looking forever if none are ever found. This behaviour reflects the fact that we have currently no procedures for deciding, given n, whether or not there are positive integers x, y, and z with $x^n + y^n = z^n$. The best one can do is "look around for such x, y, and z", and indeed that is what our flow chart does.

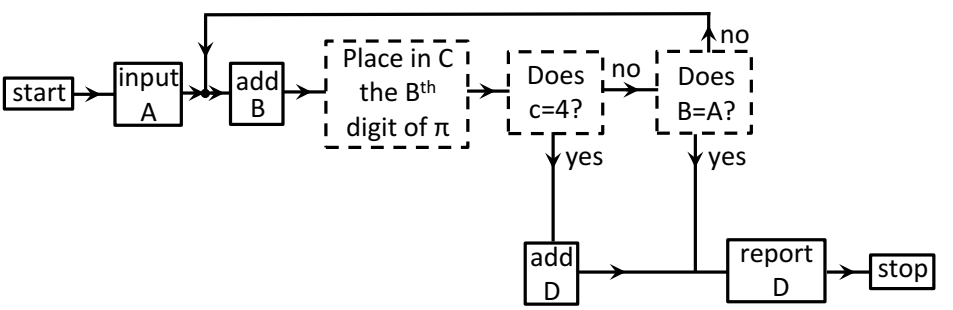
It should by this point be clear that we could continue in this way indefinitely. It should be clear that such flow charts represent precisely our intuitive understanding of the word "procedure". It should be clear that any task represented by such a flow chart would be regarded as a procedure; and any task we would regard as a procedure can be represented by a flow chart.

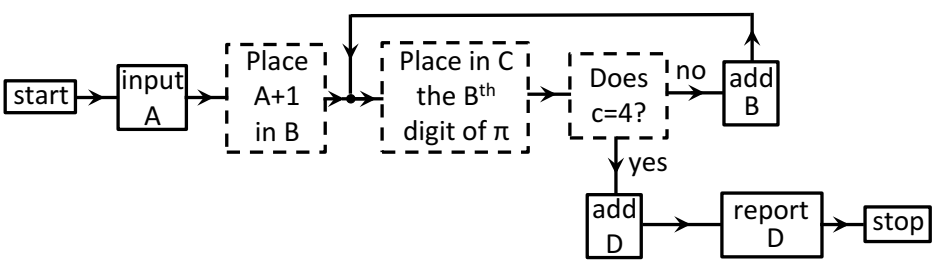

start
input A
input B
Is A less than B?
yes
no
Place A minus B in A
add C
Place A in D
report C
report D
stop
start
input A
add B
add B
Does A equal B?
no
Divide A by B, place remainder in C
C empty?
no
add B
yes
yes
stop
report D
add D
start
input A
add B
add B
yes
Does B=A?
no
Place A-B in C
Is B prime?
yes
Is C prime?
no
no
yes
add B
stop
report D
add D
start
input A
Place A+1 in B
Place B-A in C
Is B prime?
yes
Is C prime?
no
no
yes
add B
stop
report D
start
input A
input B
Divide A by 10. Place quotient in D remainder in E
remove B
B empty?
no
yes
copy D into A
stop
report E
start
input A
Divide A by 10. Place quotient in D remainder in E
add B
D empty?
no
yes
copy D into A
report B
stop

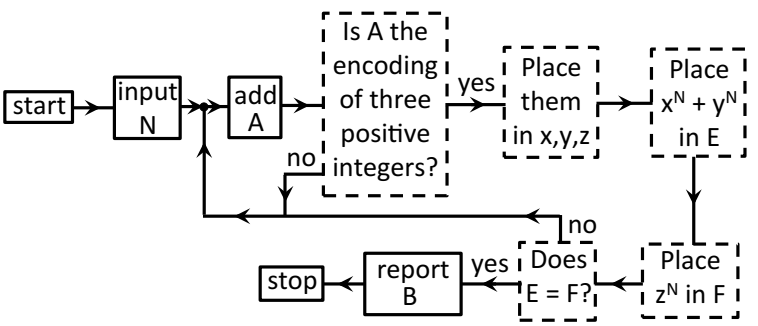

start
input A
input B
Place # digits of A in C
Place # digits of B in D
Place 2 × C + D in E
Place A × 10^E + B in F
stop
report F


start
input A
Place # digits of A in B
Divide B by 2. Place quotient in C remainder in D.
D empty?
yes
no
Are s+1st through s+cth digits of A all zero?
no
yes
no
Does s=c?
yes
stop
add s
stop
report E
report F
Place F/10^s in F
Divide A by 10^s. Place quotient in F remainder in E.


start
input A
Place in B # digits of A
Does B=q?
add c
Place in D cth digit of A
Does D=1,2,0,3?
yes
Does c=q?
no
yes
no
stop
report E
add E

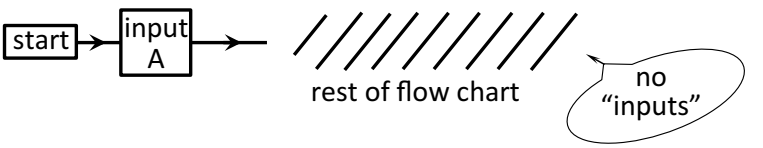

start
input A
add B
Place in C the Bth digit of π
Does c=4?
no
Does B=A?
no
yes
yes
add D
report D
stop


start
input A
Place A+1 in B
Place in C the Bth digit of π
Does c=4?
no
add B
yes
add D
report D
stop


start
input N
add A
Is A the encoding of three positive integers?
yes
Place them in x,y,z
Place $x^N + y^N$ in E
Place z^N in F
Does E = F?
yes
report B
stop
no
no

6. Will the Technician Stop?

Let us consider now flow charts of the following special form: The start "start" is to be followed in succession by "input A" and then the rest of the flow chart. This "rest of the flow chart" is to have no further inputs, and is not to access the "input A". That is, we consider flow charts of the form given below.



This form guarantees, of course, that the technician, on executing the flow chart, will request from us an input one and only one time. Note that virtually every flow chart we have considered is already in, or can easily be cast into, this form. For those having no inputs, we merely add an "input A" right after the "start", and then ignore the A-value, thus generated, in the rest of the flow chart. For those having several inputs, we merely arrange to encode all the integers to be input into the single integer A. We then input this single A-value, and have the flow chart decode.

Let us consider now the following:

Big Question: Given a flow chart of the special form above, and given the nonnegative integer we shell provide at "input A", will the technician, on executing the flow chart, stop eventually, or continue forever?

This is clearly an infinite number of questions, one for each choice of flow chart and A-value. Given the choice of flow chart and of A-value, there certainly exists an answer to our question, for the result of the technician's executing the flow chart must certainly be either to stop or to continue forever. Of course, we may not be able to figure out what that answer is. (In particular, we cannot guarantee to answer the Big Question, for given flow chart and A-value, by merely hiring a technician to execute the flow chart, for were the answer to the Big Question in that case "continue forever", the technician would never be in a position to inform us that that is the answer.) (We might remark here that the reason for introducing the "special form" for flow charts on the previous page 24 was merely to standardize what is to be input to the flow chart, in order to make the Big Question easier to state.)

The Big Question (really, questions) is about flow charts. But, as we have seen, many questions within mathematics, as well as various questions about games and other things, can be reformulated to become questions about flow charts. Consider, as examples, the following:

1. Given positive integer n, is it the difference of two prime integers?

2. Given positive integer n, do there exist positive integers x, y, and z with $x^n + y^n = z^n$?

3. Does there exist integer n ≥ 3 and positive integers x, y, and z with $x^n + y^n = z^n$?

4. Do there exist ten consecutive "7" 's in the decimal expansion of π?

5. Given positive integer n, does there appear a "4" anywhere after the n^{th} digit in the decimal expansion of π?

6. Given a tag game and initial string, will the game terminate?

7. Given the board configuration and whose move it is in five-in-a-row, is that configuration x-winnable?

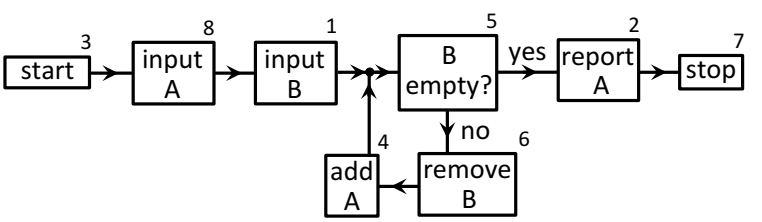
For each of these questions, one could draw a flow chart that translates our original question into the question of whether or not execution of that flow chart eventually stops. In other words, each of these questions is a special case – for some particular choice of flow chart and A-value – of the Big Question. Indeed, virtually any question in mathematics of the form "Will a systematic search for blah-blah-blah be successful?" is some special case of the Big Question.

So, we could be the envy of our tag-playing friends, capture the World five-in-a-row championship, and solve many of the heretofore unsolved problems in mathematics – if only we could answer the Big Question. But what does "answer the Big Question", as used here, mean? For any given instance – for any given choice of flow chart and A-value – one presumably *could* answer the question, merely by working on it hard enough. What is wanted here is, not the answer to the Big Question for some single instance, but rather the answer for every instance. In other words, what is wanted is a *procedure* to answer the Big Question. But what does "procedure" mean? It means a flow chart! Thus, we wish to construct a flow chart that will answer the Big Question. Such a flow chart would have to accept as inputs the "inputs" of the Big Question, namely flow charts and A-values. There is of course no problem with A-values. But in order that a flow chart accept flow charts as inputs, we must encode such charts as integers. This we now do.

Fix any flow chart. (This flow chart is to be already written in terms of the seven basic instructions, i.e., is to have all its "dashed boxes" already eliminated.) The first step is to count the total number of bowls utilized by this flow chart, and rename them using the integers from one up to this total number (in any order). The next step is to count the total number of boxes comprising this flow chart, and label them using the integers from one up to this total number (in any order). We now represent this flow chart by the following sequence of integers: number of bowls, number of boxes, integers referring to first box, integers referring to second box,. . . , integers referring to last box. Here, "integers referring to a box" are to be obtained using the following table:

start: 1, # of next box
stop: 2
input: 3, bowl #, # of next box
report: 4, bowl #, # of next box
add: 5, bowl #, # of next box
remove: 6, bowl #, # of next box
empty?: 7, bowl #, # of next box if "yes", # of next box if "no"

Thus, for example, there are three integers referring to a box containing a "remove" instruction, namely 6 (the integer code for "remove"), the number of the bowl from which the bean is to be removed, and the number of the box to which the technician is to proceed after executing this instruction. Consider, as an example, the following flow chart:



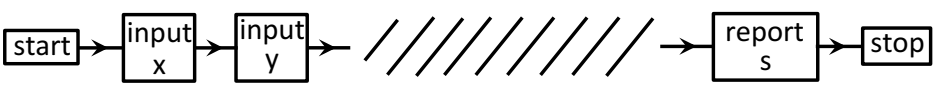
The first step is to number the bowls, say "1" for "A" and "2" for "B". The next step is to number the boxes from 1 to 8, say the numbering already given in the figure. Then this flow chart would be represented by the following sequence

of integers: 2, 8, 3, 2, 5, 4, 1, 7, 1, 8, 5, 1, 5, 7, 2, 2, 6, 6, 2, 4, 2, 3, 1, 1. Translation: There are 2 bowls and 8 boxes. Box numbered 1 is an "input" (type 3) to bowl number 2 after which one proceeds to box number 5. Box numbered 2 is a "report" (type 4) from bowl number 1 after which one proceeds to box number 7.... Box numbered 8 is an "input" (type 3) to bowl number 1, after which one proceeds to box 1.

In this way, then, we may represent any flow chart by a sequence of positive integers. We now merely encode this sequence of integers into a single positive integer using our earlier method (for an indeterminant number of integers). In this way, then, we may encode any flow chart to a single positive integer. Of course, we can also decode, to obtain the chart from the integer: Given the positive integer, we first decode, as described earlier, to obtain the corresponding sequence of integers. From this sequence, we read off in turn the number of bowls, the number of boxes, and for each box, the type of instruction it represents, the bowl (if any) to which that box refers, and the box (s, if any) to which one proceeds after executing that instruction. But this information – what the boxes contain and how they are hooked together – is precisely what is needed to reconstruct the original flow chart.

Let us return now to the issue at hand: Whether there exists a procedure (read, flow chart) to answer the Big Question. Let us call such a flow chart – whose structure and operation we are about to describe – Nancy. Now, the Big Question has two "inputs": In order to ask the question, we must select, first, some flow chart (of our special form), and, second, some A-value. Hence, the flow chart Nancy should have two input instructions, allowing us to enter, first, the integer that encodes the selected flow chart, and, second, the selected A-value. The answer to the Big Question is either "stops" or "continues forever". Hence, the flow chart Nancy should have one report instruction, at which the answer may be reported (say, using the code "0" for "stops" and "1" for "continues forever"). That is, the flow chart Nancy should have the following form:

Nancy:



The middle portion should contain only the instructions "add", "remove", and "empty?" (for we do not any starting, stopping or communicating with us to take place there). So far, we have only arranged that Nancy have the correct inputs and outputs to match those of the Big Question. Now comes the important part: We must demand that Nancy correctly answer the Big Question. In more detail, we demand:

Let the flow chart Nancy be executed, and let there be entered, at "input X" the integer that encodes any flow chart of our special form, and, at "input Y", any integer. Then the execution of flow chart Nancy must always stop, there having been reported, at "report S", "0" if that flow chart (the one whose encoding was entered at X) with that A-value (the integer that was entered at Y) stops, and "1" if it continues forever.

So far, we have merely stated what it is we want the flow chart Nancy to do. We have certainly not issued any guarantee that there exists such a flow chart. So let us now, just to get an idea of what is involved, consider how one might go about trying to construct such a flow chart. One might think first of having Nancy merely execute the flow chart (the one whose encoding was entered at X) with the A-value (entered at Y). Indeed, it is not difficult to design a flow chart that will do precisely this. But, although we may construct such a flow chart,

it would not do for Nancy. The problem is that, while it would correctly report "0" in the case of "stops", it would be unable ever to report "1" for "continues forever". Trying to construct a flow chart Nancy along these lines is analogous to trying to answer the Big Question merely by hiring a technician to execute the selected flow chart with the selected A-value.

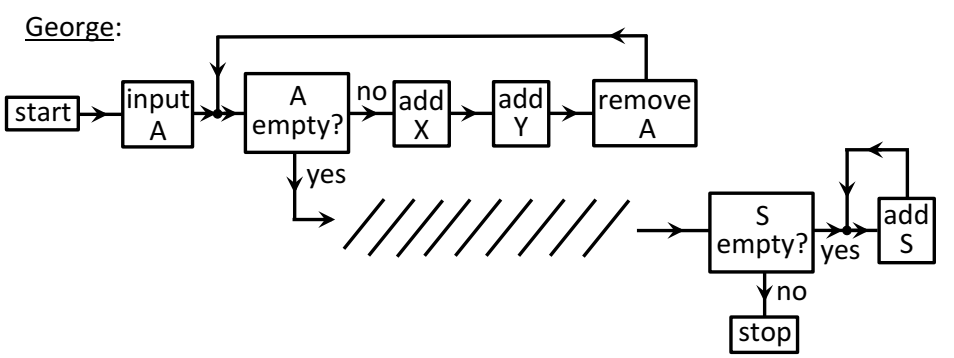
Clearly, the flow chart Nancy must, not merely execute the flow chart encoded at X, but rather examine the structure of that flow chart to determine, by some indirect argument, whether or not it (with the selected A-value) will stop. For example, the flow chart Nancy might first examine the flow chart encoded at X to determine whether or not it contains a "stop" instruction. If there is no "stop" instruction, then that flow chart, when executed, can clearly never stop. Hence, on so finding Nancy would set S equal to "1" (the code for "continues forever") and exit at the instruction "report S". Similarly, if the flow chart encoded at X contains no nodes (points where several flow lines meet), then execution of that flow chart must always stop. Hence, flow chart Nancy could check to see if the flow chart encoded at X has such nodes. If not, then Nancy would set S equal to "0" and exit. Here, then, are two examples of indirect arguments by which Nancy, on examining the structure of the flow chart encoded at X, might determine whether or not execution of that flow chart will ever stop. Clearly, much more sophisticated arguments than these two would have to be employed by Nancy. But one could in fact invent more sophisticated arguments showing that a flow chart will or will not stop, when executed, under various circumstances. Thus, it is at least conceivable that there could exist an argument so sophisticated that it could decide in *every case* (i.e., for every choice of flow chart of our special form and every choice of A-value), whether or not that flow chart, with that A-value, will stop. It is at least conceivable, in short, that there could exist a flow chart Nancy. On the other hand, one could equally well imagine that there exists no such Nancy.

Does there exist a flow chart Nancy? Is our career as a famous mathematician and world-champion five-in-a-row player just around the corner? Alas, it is not to be:

Theorem. There exists no flow chart Nancy.

Proof: Suppose for a moment that we were given a flow chart Nancy. (We shall obtain X- and Y- values for which this alleged "Nancy" in fact gives the wrong answer. From this it follows that no Nancy can exist.)

We now construct a new flow chart, which we call George, as follows:



The first three boxes of Nancy, "start", "input X", "input Y", are here replaced by six boxes, as shown; and the last two boxes of Nancy, "report S", "stop", are here replaced by three boxes as shown. But the entire reminder of the flow chart Nancy – the part indicated by / / / / / / / ; the part that does the real work – is left intact. The first three boxes of Nancy served merely to input values for X and Y. There are replaced in

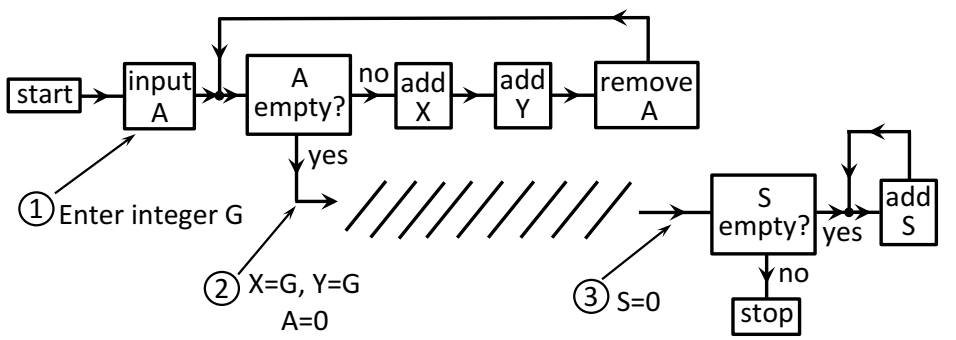
George by five boxes that input a single number A and then copy it into bowls X and Y. The last two boxes of Nancy served merely to report the value of S (either "0" or "1") and stop. These are replaced in George by three boxes that cause George to continue forever in the case S is "0", and stop otherwise. In short :

Flow chart George accepts a single nonnegative integer, copies it into bowls X and Y, and then has Nancy run with those X- and Y- values. If Nancy would have reported "0", then George continues forever; while if Nancy would would have reported "1", then George stops. But recall that reporting "0" is Nancy's way of saying that the flow chart encoded at X, with A-value at Y, stops; while reporting "1" is her way of saying that it continues forever. Thus; *flow chart George does* (by stopping or continuing forever) *the opposite of what Nancy says* (about the flow chart encoded at X, with A-value at Y).

Thus, given flow chart Nancy (and we are supposing, for the moment, that we were given a Nancy) we could perfectly well construct flow chart George as indicated. And it would do what we said it does. Now George is merely some flow chart. Hence, we may encode it as a positive integer. Call this integer G, so G is merely some specific (although very large, I am sure) integer. We may also note that flow chart George is of our special form. That is, it begins with "start", "input A", and has no further inputs.

Now comes the key part of the argument. Let us now execute flow chart Nancy, and, at both "input X" and "input Y", enter the integer G (the integer that encodes George). We claim that, whatever answer Nancy gives in this case, it must be the wrong answer.

Suppose, for example, that Nancy, when executed as above, gives value "0" at "report S". What Nancy is predicting, then, is that flow chart George (the flow chart encoded at X), with A-value G (the integer entered at Y) will stop. So let us check this prediction of Nancy by simply running flow chart George, entering integer G at "input A".



By position ② George will have completed his copying of A into X and Y. Hence, we shall at this position have G beans in each of bowls X and Y. Next, George runs Nancy with these X- and Y-values. There results, as position ③, S = 0 (for that is what we are supposing of Nancy, by the first sentence of this paragraph). Hence, George will continue forever (for he has S = 0 on entering the box "S empty?"). Thus, Nancy's prediction – that flow charge George with A-value G will stop – is wrong, for when we actually run flow chart George entering A-value G we find that George continues forever.

But what if Nancy makes the other prediction? What if Nancy, when executed with integer G entered at both "input X" and "input Y" gives value "1" at "report S"? What if Nancy predicted that flow chart George, when run with A-value G, continues forever? Again, we may check this prediction of Nancy by simply running flow chart George, entering integer G at "input A". Everything will be the same as before, until position ③, at which we shall now have S = 1. Hence,

George will stop (for he will have S = 1 on entering the box "S empty?"). So, this prediction of Nancy would be wrong too.

To summarize, we have shown that, given any candidate for a flow chart Nancy, there is some choice of X- and Y-values (namely, X = G and Y = G, where G is the integer that encodes a certain flow chart George, which, in turn, is constructed from our candidate for Nancy) for which that alleged "Nancy" must give the wrong answer. We conclude that there exists no flow chart Nancy. (End of proof.)

This is a difficult and confusing proof. In order to be fully understood, it has to be read slowly and carefully – probably on several different occasions. But the idea of the proof – stripped of all the technical details – is quite simple. We suppose that we had a candidate for our desired flow chart, Nancy. We use this Nancy to construct flow chart George. George takes a single integer (for instance, 638), and proceeds to ask Nancy "What happens if the flow chart encoded as 638 is run with A-value 638?" If Nancy answers "It stops.", then George continues forever: and if Nancy answers "It continues forever.", then George stops. That is, George *does* the opposite of what Nancy *says*. The problem arises when we use, instead of integer 638, the integer that encodes George himself, the integer we call G. For this case, George is asking Nancy "What happens if flow chart George is run with A-value G?". But, since George (by design!) does the opposite of what Nancy says, either answer by Nancy to this question must be the wrong answer. In short, our candidate for Nancy does not work. Finally, since every candidate for a Nancy does not work, there can exist no Nancy.

Note that this proof is constructive: We actually *construct*, given a candidate for a Nancy, X- and Y-values for which that candidate must fail. Imagine that one gave as a test problem "Draw a flow chart Nancy." This would of course be an easy problem to grade, for by the theorem ("There exists no flow chart Nancy.") everyone receives a score of "0" on the problem. Nut then, after the tests are graded and returned, various students come up claiming that their Nancys actually work. There students are not interested in abstract proofs about existence or nonexistence of a Nancy: They want to know what is wrong with their individual Nancys. Fortunately, the proof of the theorem is such as to make it easy for the grader to prevail in these confrontations. A given student has a candidate for a Nancy. The grader merely takes that student's candidate for Nancy, constructs the corresponding George, determines the number G that encodes that George, and says to the student: "Try your Nancy out with X-value G and Y-value G". The student's candidate for Nancy will not work for these particular values. So, that student's score, "0", is justified. Of course, different students will have different candidates for Nancy, and so will receive different numbers (G) to try their candidates on. Thus, one student might be told by the grader "Try your Nancy out with X-value 31206 and Y-value 31206." Another student's flow chart might work for *these* X- and Y-values, but that student would be told "Try your Nancy out with X-value 4110002142 and Y-value 4110002142." Each student would be given values for which his or her particular flow chart would fail.

Note that there is no "paradox" here. (I have never been able to understand what this result has to do with the familiar self-reference paradoxes, with which it is often linked. "A barber shaves everyone who does not shave himself. So who shaves the barber?") Here, we merely state a theorem, and proceed to prove it.

The construction of flow chart George, by modifying a candidate for a Nancy, is rather delicate. The modification of the beginning of Nancy serves essentially to reduce the number of "input" instructions from two (in the case of Nancy) to one (in the case of George). The reason for this reduction is the following. Nancy deals only with flow charts "of our special form" (i.e., having "start", "input A", and no further inputs). But Nancy herself is not of this form, for she has two inputs.Thus, as things stand, Nancy cannot ask about herself. Flow chart George, on the other

hand, is "practically Nancy", but, because of the modification of inputs, *is* of our special form. Thus, Nancy can ask about George, and so, indirectly, about Nancy herself. It is asking Nancy, indirectly, through George, about herself that makes the proof work.

Why are the other details of George as they are? Why does George copy the *same* integer into both X and Y? Why does George *do* the opposite of what Nancy says, as opposed, e.g., to *saying* the opposite of what Nancy says? Or, why not have George do the *same* as Nancy says? The easy answer to these questions is "if you construct George differently according to one of these proposals, then you do not get a proof of the theorem." After all, a proof in mathematics has only the obligation to be a proof. It does not have to be transparent, natural, or even interesting. I must confess that I do not have any deeper answers to these questions. The construction works.

We have agreed to take "procedure" to mean "flow chart". Thus, the meaning of the theorem is that there exists no procedure to determine, for any given flow chart of our special form and any given A-value, whether that flow chart with that A-value will stop or not. Imagine that it were your job to receive each day a flow chart (of our special form) and A-value and to determine whether that flow chart with that A-value will stop. Some would be easy; some would be hard. But, presumably, one could in every case – given sufficient effort and ingenuity – find the answer. But effort and ingenuity are hard work. One might look for an easier way. One might look for some procedure that could be followed in every case – some procedure that would be guaranteed in every case to produce the answer. The theorem guarantees that no such procedure can ever be found. There is, if you like, always a place for effort and ingenuity.

This situation should be contrasted with that of problems we have discussed earlier.

We have given examples of questions for which there is a procedure (Given positive integer n, is n prime?), and of questions for which we do not know whether or not there is a procedure (Given positive integer n, do there exist positive integers x, y, and z with $x^n + y^n = z^n$?). But we had not given earlier any example for which one could positively assert that there *is no* procedure. Here, from that theorem, we have our first such example.

There are more levels, than one might have expected naively, at which one can understand a question. No longer are there merely questions for which "We know that we can determine the answer." and those for which"We do not know whether or not we can determine the answer." Now, there are in addition questions for which "We know that we cannot determine the answer."

7. Appendix

Phy Sci 113 Sep 27, 1988
Due: Oct 4, 1988

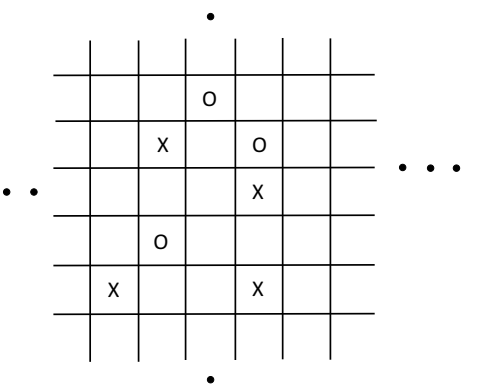
Problem Set 1

1. Learn to play tic-tac-toe expertly. (By "expertly" I mean such that yoou can always make the best possible move. Most of you probably know how to do this already. If you do not know the game, please ask, a friend.) Explain, in no more than a page, the procedure for expert play.

2. Nim is a game for two players. A number of beans is first counted out into a bowl. The players alternate turns, with each player, in his turn, taking either one or two beans from the bowl. The player taking the last beans wins. Learn to play nim expertly. Explain the strategy. (It is easier than it sounds! First try some games with a small number of beans.)

3. Learn to play checkers passably. (By "passably" I mean knowing the rules, and playing at a level of competence that comes from having played a few games. We shall adopt in checkers the following two rules: i) You must make a jump if one is available; and ii) if, during the course of the game, the board returns to exactly the same position for the third time, then that game is declared a draw.)

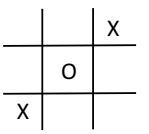
4. Five-in-a-row is a game played on a grid of vertical and horizontal lines. (The grid is "infinite", i.e., additional lines on the sides and top and bottom can be added as needed.) There are two players, who alternate turns. Each player, in his turn, makes his mark (say, one player "x", one "o") on any unoccupied square. The first player to get five of his marks consecutively in a row (horizontally, vertically, or diagonally) wins. Learn to play five-in-a-row passably.



Phy Sci 113 Oct 4, 1988

Problem Set 1 — Solutions

1.
Rule 1. If you are able to win, i.e., to achieve three in a row, on that turn, do so.
Rule 2. If you are able to block your opponent's win, i.e., to take the third in a row in which he has two, do so.
Rule 3. If the center is available, take it.
Rule 4. If you hold the center only, and your opponent holds just two opposite corners, take one of the sides.



O = you

Rule 5. If a corner is available, take it.
Rule 6. Take any move.

The strategy is to move according to Rule 1 if you can. If you cannot then try to move according to Rule 2; if you cannot, then according to Rule 3; etc. Continue down the rules until you are able to make a move.

2. It is your move. Take the number of beans then in the bowl, divide by three, and find the remainder. (Thus, for example, if there are eleven beans in the bowl, then the remainder is two; if there are twenty-one, then the remainder is zero.)

If the remainder is zero, then take either one or two beans from the bowl.
If the remainder is one, then take one bean from the bowl.
If the remainder is two, then take two beans from the bowl.

In the first case, you will lose if your opponent plays expertly.
In the last two cases, you are assured of a win by this strategy.

Phy Sci 113 Oct 4, 1988
Due: Oct 11, 1988

Problem Set 2

1. Consider the tag game with just two symbols, a and b, and with the following rules:

A. If the leftmost symbol of the string is "a", then append "aa" to the right and remove the leftmost three symbols.

B. If the leftmost symbol of the string is "b", then append "bbab" to the right and remove the leftmost three symbols.

a. Show that if the initial string is of the form "axxaxxaxx... axx", where each "x" stands for either an "a" or a "b", then the tag game eventually terminates at the empty string.

b. Show that if the initial string is of the form axxbxxaxxbxx... axxbxx", then the tag game eventually cycles.

c. Try to find an initial string for which the tag game goes on indefinitely, without terminating or cycling. (Note: This may be difficult. I tried it for an hour or so, without success. So, if you have not found one after reasonable effort, just describe for your answer what methods you tried, and how they failed.)

2. Consider the tag game above, but with "a" leftmost append, instead of "aa", and "aba"; and with "b" leftmost append, instead of "bbab", "bbba". For this game, give the procedure for deciding, for any given initial string, whether the game eventually terminates at the empty string, cycles, or goes on indefinitely without terminating or cycling.

3. Select any two board games with which you are familiar, other than those discussed in class. For each of your games, discuss whether the problem of making expert moves is well-posed, whether it has a solution, and whether there is a procedure for expert play.

Remark: If, for the tag game of problem 1, you can figure out what happens with the initial string "bbbbbbbbbbbbbbbbbbbbb" then you will become famous, for this is an old, much-worked-on, and as yet unsolved problem.

Phy Sci 113 Oct 11, 1988

Problem Set 2 — Solutions

1.a. Suppose, e.g., that the initial string were "axxaxxaxxaxxaxx". Then the results of the first five moves would be:

axxaxxaxxaxxaa, axxaxxaxxaaaa, axxaxxaaaaaa,
axxaaaaaaaa, aaaaaaaaaa.

Thus, after five moves the string is all "a" 's. But now each further move will merely add two "a" 's to the right while deleting three from the left, i.e., each further move will again result in a string with all "a" 's, but with one fewer "a" for each move. Thus, successive moves from here will produce: aaaaaaaaa, aaaaaaaa, aaaaaaa, aaaaaa, aaaaa, aaaa, aaa, aa, a, (empty). Similarly, an initial string with n "axx" 's will, after n moves, be all "a" 's, after which it will shrink to the empty string.

b. The result of the first two moves on the string "axxbxxaxxbxx... axxbxx" are

bxxaxxbxx... axxbxxaa, axxbxx... axxbxxaabbab.

Thus, we obtain after two moves a string of exactly the same form as the original string, but with the "x" 's in the final six entries made explicit. (That is, instead of ending in "axxbxx", our string now ends in "aabbab".) Repeating this a number of times equal to the number of "axxbxx" 's in our original string, we end up with the string "aabbabaabbab... aabbab" (i.e., the string with each original "axxbxx" replaced by "aabbab"). But now two further moves on this string give the same string back. So, we have arrived at a cycle.

c. I still have not found such an example, and I am beginning to think that I am not going to find one soon. Here is just one possible strategy (which seems to fail) to search for an example. Suppose that one could find a string such that, after one has made enough moves to "use up" (i.e., totally delete) that string, what results is the original string again, but with some additional letters on the right. One might then hope that, as the game continues, one will again recover the original string, but with still more letters on the right, etc. Then, with any luck, one will be able to guarantee that the lengths of the string will increase without bound. But it seems difficult to implement this strategy. Here is a good-looking candidate: bxxbxxbx. Here are the results of the first three moves: bxxbxbbab, bxbbabbbab, babbbabbbab. Note that we now have the form of the original string ("bxxbxxbx", here made explicit, to "babbbabb") with some additional letters (namely, "bab") on the right. It looks promising. But let us continue the game from here: bbabbbabbbab, bbbabbbabbbab, abbbabbbabbbab, babbbabbbabaa. It looks very promising at this point! We once again have our original string, but now with "babaa" on the right. But if we continue to play from here, things break down: bbabbbabaabbab, bbbabaabbabbbab, abaabbabbbabbbab, abbabbbabbbabaa, abbbabbbabbaaaa, babbbabbaaaaaa. I just do not see how, from here, to guarantee that we will continue, after every few moves, to recover our original string, plus more and more extra letters on the right.

2. First note that, under these rules, no game ever terminate at the empty string. Indeed, each step in the game either leaves the length of the string the same (in the case the leftmost letter is "a"), or increases the length of the string by one (leftmost letter "b"). Thus, we have only to decide whether the game cycles, or goes on to ever longer strings.

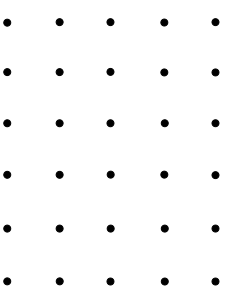
If, during the course of a game, a "b" were ever hit (i.e., were the leftmost letter) then we would append "bbba" on the right. But these three "b" 's in a row assure that one of them would be hit eventually (since, at each step, just three are removed from the left). This would result in three more "b" 's on the right, and so in another "b" hit, etc. We conclude: If, in the course of a game, a"b" is

ever hit, then the game will go on to ever longer strings (since, as we have just seen, we must then continue to hit "b" 's). But clearly, if no "b" is ever hit, then the game must cycle (for then only "a" 's would be hit, and so the string length must remain the same).

Thus, there are just three candidates for games that cycle (i.e., those for which only "a" 's are hit), namely those with initial strings of the form "axxaxx... axx", "axxaxx... axxax", "axxaxx... axxa". Playing each of these initial strings for one step, we obtain, respectively, "axx... axxaba", "axx... axxaxaba, and "axx... axxaaba". We see that the first and third, but not the second, continue to be of our form. We thus obtain:

Answer: All initial strings of the form "axxaxx... axx" or "axxaxx... axxa" lead to games that cycle. All others continue to ever longer strings. No game terminates.

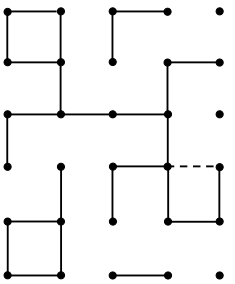
3. For the game of bridge, the problem is not well-posed in our sense. In the first place, there is concealed information (in that the hands are not shown to others), and so there is no board configurations giving the entire information about the status of the game. In this sense, bridge is more like poker. A further complication is that, in bridge, there are four players, acting in pairs as teams. Thus, a "player" in our sense would be an entire team, and one would have to allow the "players" to have two stores of information (those possessed by the two members of that team), such that information could not readily be passed back and fourth. Because of these complications, it is not even clear (to me, at least) that there exists "expert play" in bridge using probabilities for various moves (as there is, e.g., in poker). In any case, bridge is very far from the types of games we have in mind.



Consider, as a second example, the pencil-and-paper game "boxes". One begins with a finite grid of points on a piece of paper. There are two players, and each in his turn is allowed to connect two adjacent points by a horizontal or vertical line. Thus, the figure on the next page illustrates a typical board configuration. The dashed line indicates a move whose turn it is.

If, in your move, you complete the fourth side of a "box" (one of the little squares having four of the points as corners), then you are awarded one point, and you move again. Thus, in the example at the right, the person about to make the dashed-line move is about to receive one point (and to take another move). At the end (when all lines have been drawn in), the player with the most points wins the game.

This game, I claim, satisfies all of our criteria. There is a board with configurations giving all information about the game. There are two players, who alternate turns, and make certain allowed moves. There is a win, lose, and draw. (To make this clearer, one should have the players mark (with an "x" or "o") squares they have completed. Then the "board configurations", would include these marks on the squares. The "x-win configurations", then, would be those in which the entire board is filled in, and "x" has more squares than "o".



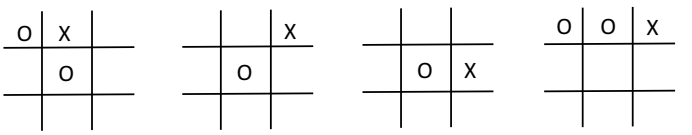
We conclude that, for the game of boxes, the problem of expert play is well-posed, and there does exist an expert strategy. It is not clear whether or not there exists a procedure for expert play. (Certainly there does if the size of the initial grid is fixed, for then there are but a finite number of board configurations. But for arbitrary games, there is no finite number.)

Do you think there is a procedure for expert play? What is your guess?

Phy Sci 113 Oct 11, 1988
Due: Oct 18, 1988

Problem Set 3

1. In each of the tic-tac-toe configurations below,



it is x's move. For each, decide whether that configuration is x-winnable, o-winnable, or drawable, and give a good move for x.

2. Show that, for a game within our framework, if it is x's turn and the board configuration is drawable, then there exists a move for x that results in a drawable board configuration.

3. Given a positive integer n,

a. is it the sum of any number of prime numbers?
b. is it the square of a prime number?
c. is it any power of a prime number?
d. is it the sum of two odd integers?
e. is it the difference of two odd integers?

For each of these questions, determine whether or not there is a procedure for answering the question. If there is a procedure, give it.

4. Given positive integer n, give explicitly the procedure for deciding whether or not it represents an "x-wins" configuration in tic-tac-toe. (Make sure that your procedure really is that – something that can be followed mechanically.)

Phy Sci 113 Oct 19, 1988

Problem Set 3 — Solutions

1. Label the squares on a tic-tac-toe board as shown below.

1	2	3
4	5	6
7	8	9

a. This configuration is o-winnable. Thus, every move for x results in an o-winnable configuration, and so every move for x is a good move.

b. This configuration is drawable. It turns out that every possible move for x results in a drawable configuration, and so every move for x is a good move.

c. This configuration is drawbale. A move by x to 4 results in an o-winnable configuration, while all other moves – to 1, 2, 3, 7, 8, or 9 – result in a drawable configurations. Hence, any of these six moves is a good move.

d. This configuration is x-winnable. Moves to 6 or 9 result in an x-winnable configuration; to 5 or 8 in a drawable configuration; and to 4 and 7 in an o-winnable configuration. Hence, the good moves are to either 6 or 9.

2. The board configuration is drawable, and it is x's turn. There can be no move for x that results in an x-winnable configuration, for were there such a move then, by the rules that define x-winnable, the original configuration would already have been x-winnable. So, every move for x results either in a drawable or an o-winnable configuration. Were it the case that every move results in an o-winnable configuration, then, by the rules that define o-winnable, the original configuration would already have been o-winnable. Hence, there must be some move for x that results in a drawable configuration.

3.a. There does exist a procedure: "If the integer is 1, say 'no'; otherwise say 'yes'." (Every integer 2 or greater is the sum of some number of prime integers. Indeed, every such integer is the sum of some number of 2's and 3's.)

b. There does exist a procedure. Suitable instructions might direct that the technician take in turn each integer from 2 up to the given integer. First, test to see if it is prime, and then test to see if its square is the given integer. If in any case both tests are positive, say 'yes'; if none is found by the time you reach the given integer, say 'no'.

c. There does exist a procedure. Let the given integer be n. For each integer from 2 up to the given integer, first test to see if it is prime. Then begin raising that integer to various powers – first, second, and so on, continuing until the result exceeds n. Test to see whether the result ever equals n. If in any case both tests are positive, say 'yes'; if none is found by the time you reach n, say 'no'.

d. There does exist a procedure. Say 'yes' if the integer is even, and 'no' if it is not.

e. There does exist a procedure. Say 'yes' if the integer is even, and 'no' if it is not.

4. Say 'yes' if the given integer is one of the forms lxxlxxlxx, xlxxlxxlx, xxlxxlxxl, lllxxxxxx, xxxlllxxx, xxxxxxlll, lxxxlxxxl, xxlxlxlxx; and 'no' otherwise. Here, "x" can stand for any digit. (Note that we have, above, merely listed all possibilities for three x's in a row.) Remark: There is of course also a procedure to decide whether a given configuration is x-winnable, but that is far more

complicated than this one. Basically, one would have to have the technician reconstruct the tic-tac-toe board, then play out the game using our earlier expert strategy, find out who wins, and answer accordingly.

Phy Sci 113 Oct 18, 1988
Due: Oct 25, 1988

Problem Set 4

1. Given positive integer n, is it the encoding of three positive integers, x, y, and z, such that x + y = z? Answer this question for each of the following n:

a. 30000000150008

b. 13000000000000000200000000213

c. 1300000000000000200000000213

2. Is there a procedure for answering?

a. Given positive integer n, do ten 7's occur in a row in the first n digits in the decimal expansion of π?

b. Given positive integer n, is both n a prime and there exist positive integers x, y, and with $x^n + y^n = z^n$?

3. Encode the four positive integers 13, 2, 11, 100 into a single integer.

Phy Sci 113 Oct 25, 1988

Problem Set 4 — Solutions

1. a. No, The given integer is indeed the encoding of 3, 15, and 8. But it is not the case that $3 + 15 = 8$.

b. Yes. The given integer is the encoding of 13, 200, and 213. It is the case that $13 + 200 = 213$.

c. No. The given integer is not the encoding of any three positive integers. Indeed, it has an odd number (27) of digits.

2. a. There is a procedure. Use any of the known procedures for computing the first n digits of π, then inspect those digits to see if anywhere there occur ten "7" 's in a row.

b. We do not know whether or not there is a procedure. If the given n is not a prime, then clearly the answer is "no". But what if n is a prime? Then we do not have, offhand, any way to tell whether or not there exist x, y, and z with $x^n + y^n = z^n$. Thus, at least as the things stand, we have no procedure.

3. Encoding 11 and 100 yields 1100000100. Encoding 2 and this yields 200000000001100000100. Encoding 13 and this yields 13000000000000000000000000200000000000110
This last monstrosity is the desired answer.

Phy Sci 113 Oct 25, 1988
Due: Nov 1, 1988

Problem Set 5

1. To specify a tag game, one must tell what is the alphabet, what is to be appended on the right according to which letter is on the left, how many leftmost letters are to be deleted, and what is the initial string. Encode all this information in a single positive integer.

2. Given positive integer n, do there occur n or more consecutive "7" 's anywhere in the decimal expansion of π? Is there a procedure for answering this question? (Note: This is a *tricky* problem. It is worth thinking about for a while.)

Phy Sci 113 Nov 1, 1988
Due: Nov 8, 1988

Problem Set 5 — Solutions

1. There are of course many ways of encoding this information. Here is one. Given the tag game and initial string, first represent it as a sequence of integers, as follows:
total number of letters in the alphabet (we represent these letters by the numbers from 1 up to this total), number of letters in the string to be appended on the right if the first letter appears on the left, the successive letters (represented as numbers) in this string, number of letters in the string to be appended on the right if the second letter appears on the left, the successive letters in this string,... (up to these things for the last letter in the alphabet), number of letters from the left to be deleted in each turn, letters (represented as numbers) in the initial string.
Thus, for example, the tag game of the first test, with initial string "abbaa", would be represented by the following sequence of numbers: 2, 2, 2, 2, 3, 1, 1, 1, 3, 1, 2, 2, 1, 1. Translation: There are 2 letters in the alphabet. If the first (a) appears on the right, append the 2-letter string consisting of "bb" (2 and 2). If the second (b) appears on the right, append the 3-letter string consisting of "aaa" (1 and 1 and 1). Remove each time 3 letters from the left. The initial string is "abbaa" (1 and 2 and 2 and 1 and 1).

Thus, we have so far represented a tag game, with initial string, as some finite sequence of integers. We now merely encode these into a single integer as was shown in class. (Here, of course, we must use "encoding an indeterminate number of integers", because we do not know beforehand how many there will be.) It is clear that the final integer can be decoded to find the tag rules and initial string.

2. (*Note:* I only put this problem on because it reflects a point I wanted to make, but for which I did not wish to use more class time. I did not really expect anyone to get the answer, below, but rather wanted to get you thinking about it in preparation for this point.)

There does exist a procedure to answer this question! To see this, let us consider a few possibilities.

Suppose that there are sequences of consecutive "7" 's of arbitrary length in the expansion. (This means, in more detail, that given any positive integer there is some sequence of consecutive "7" 's in the expansion of length larger than this integer.) Given this supposition, the answer to our question is "yes" no matter what n is (for there are such sequence longer than any given n). So, in this case there does indeed exist a procedure to answer the question, namely "say yes".

Suppose that the longest sequence of consecutive "7" 's in the expansion consists, say, of 188 "7" 's. In this case, there is also a procedure to answer the question, namely "Say 'yes' if n is less than or equal to 188; and 'no' if n is 189 or greater."

Suppose the longest sequence of consecutive "7" 's in the expansion consists of just 5 "7" 's. Then there is again a procedure: "Say 'yes' if n is less than or equal to 5; and 'no' if n is 6 or greater."

To summarize, then, there is in every case a procedure. If there is no longest sequence of consecutive "7" 's, then the first case above applies, with that procedure; if there is a longest sequence, then one of the other cases applies. We conclude: There does exist a procedure to answer this question.

But this is another example in which, although we know that a procedure *exists* (here, because we can list a bunch of procedures and be sure that *one* of them is the right one), we do not know which procedure is right. It is quite similar to those examples of "yes-or-no" questions in class, in which we knew that one

of two procedures ("say 'yes' " and "say 'no' ") was correct, but not which. The only difference is that, here there are an infinite number of procedures in our list. Again, one is correct, but we do not know which one.

This example very well illustrates the point that "not being able to find a procedure offhand" is quite different from "there is no procedure". Here, it turns out that there is a procedure, but it is just somewhat subtle.

Phy Sci 113 Nov 1, 1988
Due: Nov 8, 1988

Problem Set 6

Note: There are three skills to be acquired: i) how, given a simple procedure, to write a flow chart for it, ii) how, given a more complicated procedure, to explain clearly how one could, at least in principle, write a flow chart for it, and iii) how, given a flow chart, to figure out what procedure it performs. These skills can be mastered, in my opinion, only by playing around with these things on your own. The problems below are intended only to get you started in this.

1. Write a flow chart that accepts two nonnegative integers A and B, divides A by B (provided B is not zero), and reports the quotient and remainder.

2. Write a flow chart that accept two nonnegative integers A and C, and reports A – C if A is greater than or equal to C, and C – A otherwise.

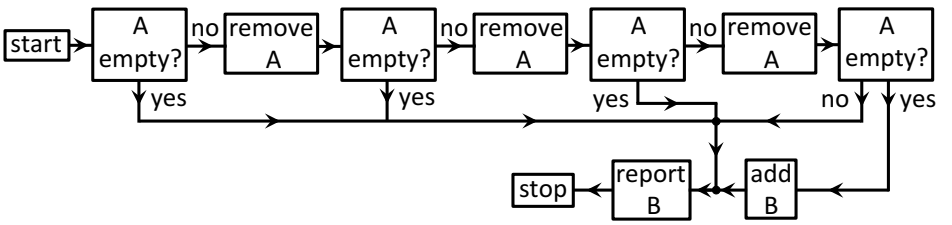
3. Write a flow chart that accepts three nonnegative integers X, Y, and Z, and reports the largest.

4. Argue that there exists a flow chart that accepts a nonnegative integer A and reports the number of digits in A.

5. Argue that there exists a flow chart that accepts a nonnegative integer A and reports whether or not A is the encoding of two other positive integers.

6. Argue that there exists a flow chart that accepts positive integers P and Q, and reports the result of our encoding of these two integers.

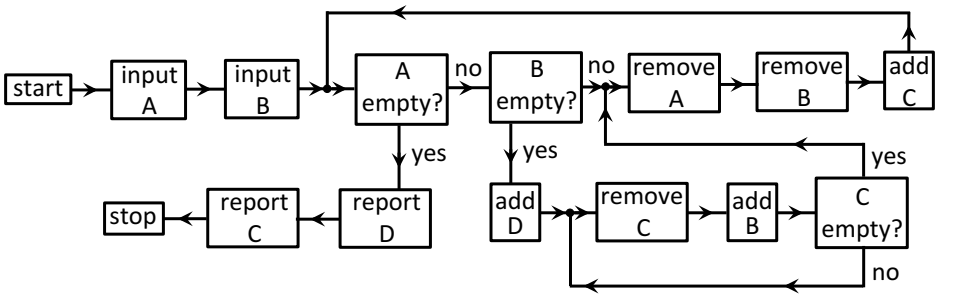
7. What does the flow chart below do?


start
A empty?
no
remove A
A empty?
no
remove A
A empty?
no
remove A
A empty?
yes
yes
yes
no
yes
stop
report B
add B

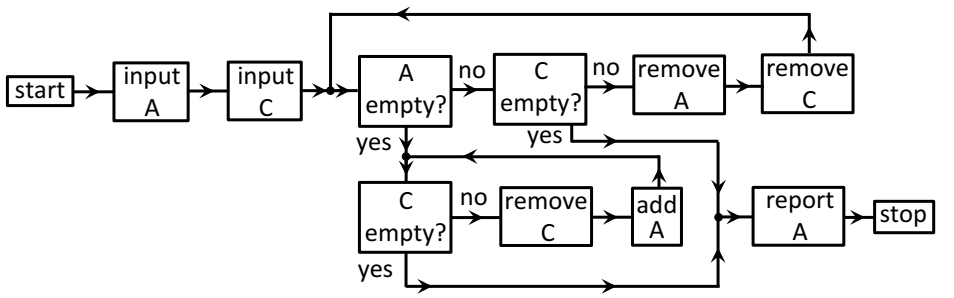
Phy Sci 113 Nov 8, 1988

Problem Set 6 — Solutions

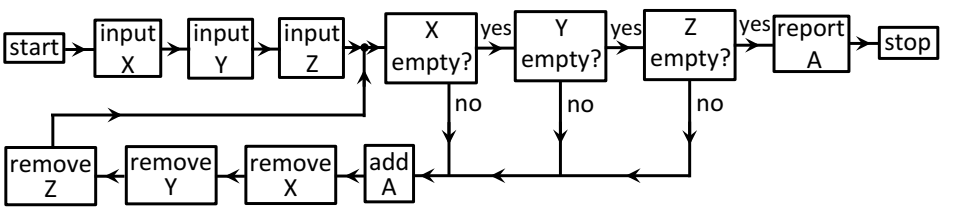
1.



2.



3.



4, 5, 6. These three problems were done in class, and also are done in the notes.

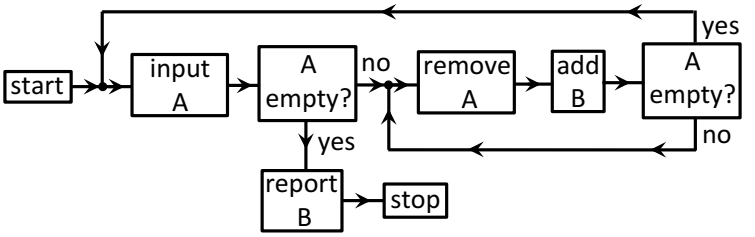
7. As the flow chart stands, the answer is "Always report zero." Correcting the flow chart by inserting the extra box "input A" after the first box (as I asked you, in class, to do), the flow chart answers the questions "Is the number A three, or not?" The result is reported as "1" if the answer is yes, and"0" if no.

Phy Sci 113 Nov 8, 1988
Due: Nov 15, 1988

Problem Set 7

1. Draw a flow chart that accepts as input three nonnegative integers, and reports which of these (first, second, or third) is the largest.

2. What does the flow chart below do?



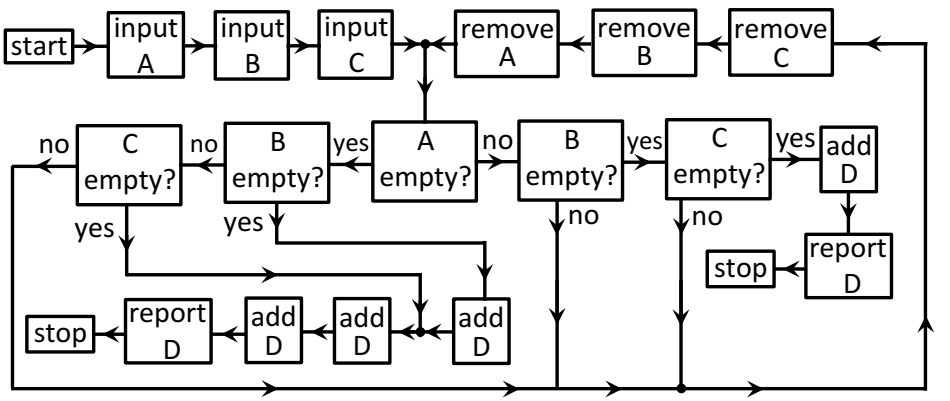
3. Draw a flow chart that accepts as input a nonnegative integer, and reports whether or not that integer is the largest prime.

4. Argue that there exists a flow chart that accepts as input a nonnegative integer, and reports whether or not that integer is equal to the sum of the squares of two integers.

Phy Sci 113 Nov 15, 1988

Problem Set 7 — Solutions

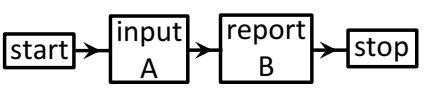
1.



This flow chart (which admittedly, is not very elegant) works provided there *is* a largest (as opposed to two- or three-way tie for largest).

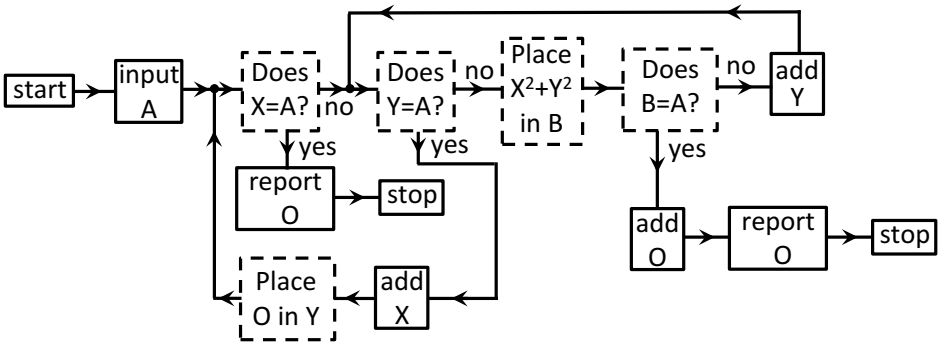
2. The user is continually asked to input numbers (a running total of which is kept in bowl B). As soon as "0" is input, this total is reported and the execution stops. In short, this flow chart is an adding machine?, with "0" the signal for for "give me the sum".

3.



Here, "0" is reported for "is not the largest prime", and "1" for "is the largest prime".

4.



Here, a “0” is reported for “no”, and “1” for “yes”.

Phy Sci 113

Nov 15, 1988
Due: Nov 22, 1988

Problem Set 8

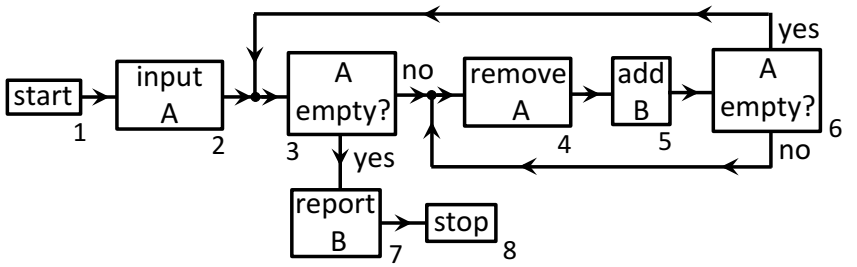
1. Translate the flow chart of problem 2, Set 7 into a sequence of integers. If this were to be encoded into a single integer, approximately how many digits would that integer contain?

2. Draw the flow chart represented by the following sequence of integer: 3, 7, 5, 2, 5, 2, 1, 4, 3, 1, 7, 5, 3, 7, 6, 1, 1, 7, 1, 2, 6.

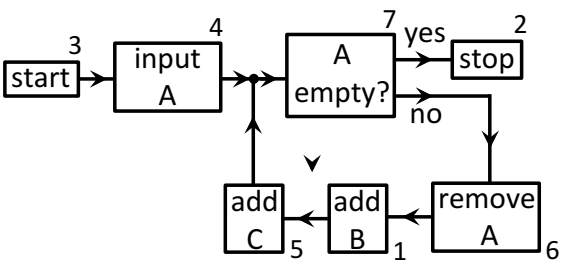
Phy Sci 113 Nov 22, 1988

Problem Set 8 — Solutions

1. 2, 8, 1, 2, 3, 1, 3, 7, 1, 7, 4, 6, 1, 5, 5, 2, 6, 7, 1, 2, 4, 4, 2, 8, 2. Here, we are numbering the bowls with "1" for "A", and "2" for "B". We have numbered the boxes as shown below:



2.



The bowls "A", "B" and "C" are numbered "1", "2", and "3", respectively. This flow chart, by the way, is from page 16 of the notes.

1. (continued) Each time we include one more number in our encoding (to a single integer), we approximately double the number of digits (actually, a bit more than double). Here, there are 26 numbers to be encoded (including the first number, "25", the number of numbers). Hence, there will be over 2^{26} digits. This is more that 60,000,000 digits!

Phy Sci 113 Oct 25, 1988

Test 1

Please write your name on every page. Please answer and *explain* your answer *briefly*.

Name:

1. (20 points)

a. A tic-tac-toe game has just begun, and it is x's turn to make the first move. Is the board configuration x-winnable, drawable, or o-winnable? Explain briefly.

b. A game within our general framework is underway, and the board is in some configuration. It is o's turn, and there are exactly eleven moves available to o. Six will result in an x-winnable configuration, three in a drawable configuration, and two in an o-winnable configuration. Is the present configuration of the board x-winnable, drawable, or o-winnable, or is there insufficient information to tell? Explain *briefly*.

2. (16 points) Give any prescription to encode an arbitrary integer (i.e., which may be positive or negative) as a positive integer.

3. (16 points) Consider the tag game with "a" and "b", and with the following rules: For "a" leftmost, append "bb" on the right, and for "b" leftmost append "aaa" on the right; and in both cases delete the leftmost three letters. Give a procedure to determine, given the initial string, whether the game will ultimately terminate, cycle, or go on forever without terminating or cycling.

4. (30 points) You are to be given positive integer n. For each of the question below, state whether there exists a procedure to answer the question, there does not exist such a procedure, or you do not know whether or not there is a procedure. Explain *briefly.*

a. Has there occurred a "4" by the n^{th} digit in the decimal expansion of π?

b. Does there occur a "4" anyplace after the n^{th} digit in the decimal expansion of π?

c. Is n either the result of encoding two integers, or the difference of two primes?

Phy Sci 113 Oct 25, 1988

Test 1 — Solutions

1.a. The configuration is drawable. That is, it is not possible for x, absent a blunder by o, to be assured of a win; nor is it possible for o, absent a blunder by x. In other words, with expert play on both sided the game will certainly be a draw.

b. The configuration is o-winnable. Recall that if it is o's turn, and there is *any* move for o that results in an o-winnable configuration, then the present configuration is o-winnable. But that is the situation here.

2. There are many such prescriptions. For example, write out the digits of the integer, and affix a "1" on the right if it is positive, and a "2" if it is negative. Or, as a second prescription, ignore the sign the integer, double it, and then add one if the original integer was negative.

3. The procedure is: "Say, no matter what is the given initial string, 'It will terminate'." Clearly, every game will either cycle or terminate (since one adds at most three letters to the right, while deleting three, and so the string-length can never grow). The only possibility for cycling, then, would be if only "b" 's are hit. But a "b" hit results in appending "aaa", which in turn assures that eventually an "a" will be hit. Hence, the string must ultimately shrink to nothing.

4. a. There is a procedure: "Say 'no' if n is one, and 'yes' otherwise." This works because the digit of π ($= 3.14159\ldots$) is a "4", and so the answer is 'yes' for n two or greater, but, clearly, 'no' for $n = 1$.

b. We do not know whether or not there exists a procedure. There is no obvious way, given n, to find the answer. Thus, for example, one could not simply try out every digit past the n^{th} in the decimal expansion, for there are an infinite number of such digits.

c. We do not know whether or not there exists a procedure. Of course, if n is the result of encoding two integers, we say 'yes'. But what if n is not? Then we know offhand of no way, by some finite procedure, to determine whether or not n is the difference of two primes.

Phy Sci 113 Oct 27, 1988

Test 1 — Comments

This test was too easy for (or, if you prefer the other way of saying it, the depth of your understanding exceeded my already very high expectations). It is not that the average (56, out of a maximum 82) was so high – although it was somewhat higher than I expected – but rather that there were a fair number of people who were just not challenged by the test. This is clear from your comments, and also from the grade distribution:

Score	Nmbr receiving	Score	Numbr receiving
10-19	1	50-59	31
20-29	6	60-69	46
30-39	9	70-79	30
40-49	24	80-89	11

What I should have done here – and will try to do next time – is including at least one very hard problem, so those who are having no trouble with the material will have something to chew on. Also, this problem should in part take care of itself, for we are now, moving into somewhat meatier material.

Comments on individual problems:

1.a. This was intended as a "throw-in" problem, to get people relaxed, and, indeed, very few had much trouble with it. A few did not realize that either player in tic-tac-toe can, from the beginning, force a draw. A board configuration must be *exactly one* of x-winnable, o-winnable, or drawable – it cannot be two or more of these.

b. The status – x-winnable, o-winnable, drawable – is a property *only* of the board configuration (and whose turn it is). It does not depend on whether people play expertly, badly, or not at all. (Of course, the definitions of these terms use the rules of the game, i.e., they envision players, etc. But ultimately it is just the configuration itself that is one of these three.) In particular, o-winnable requires only that there *exists* a move for o that makes the configuration o-winnable. It is not necessary that "the game be allowed to continue, with o actually making such a move". Nor is it necessary that o be an expert player, make good moves, or indeed even be present. If you like, these terms refer to potentialities, not actualities.

2. Most people (including whoever it was who wrote the solution sheet) forgot that "0" is also a legal integer, and so must be encoded into something. Thus, for example, both solutions on the solution sheet are wrong. (The first could be correct by the change "... if it is negative *or zero*."; the second by the change "... original integer was negative *or zero*.") (Recall that zero is neither positive nor negative.) This minor issue was assigned one point. All that counts about an integer is its *numerical value*, not the details of how it is written. Thus, one can write the integer "47" as "047" or "0047" or "+47" or "47.000" or "47" written with orange ink, but they all count as the same integer. In particular, then, one cannot encode additional information into the integer by written it in one of these alternative forms. The key thing about encoding is not only that it can be carried out (in this example, to obtain a positive integer from any given integer), but that it can also be decoded (in the example, to recover the given integer from the positive integer). Thus, for example, "Just drop the sign." and "Square the integer." are not legitimate encodings, for neither can be decoded. Thus, for the

first, if I give you the result is "7" you do not know whether the original integer was "7" or "-7"; and similarly for the second. Here you were, of course, to encode a *single* integer into a positive integer (as opposed to several integers, before in class). (As you study this material, try to focus on what is really going on ("the idea of encoding") rather than the mechanics ("how to encode").) The class as a whole did rather well on this problem (which I thought would be more difficult). There were some very original encodings.

3. Please *explain your answer* briefly. There were a fair number of papers with "Just say 'terminates'." and no further explanation (these assigned 10 out of 16 points). It was also common to figure out, by some means, that every initial string would terminate, but then forget to state what the procedure is ("Just say 'terminates'.") (I assumed that people who had figured all this out knew also what the procedure is.) While most explanations of why every string must lead to termination were quite clear, a number were less so. What of the answer "Let the technician play out the tag time, starting from the given initial string, and see what happens."? Of course, this is general not a procedure, for it is conceivable that the technician could go on playing the game forever, without ever being able to announce "This game goes on forever". It is curious, however, that *for particular example of the problem*, this actually is a procedure. The reason is that it is in fact the case that every game will terminate, and so the labors of the technician will also come to an end, with that technician announcing "terminates". This is the old story: It need only be the case that the technician's efforts terminate, not that the technician is aware of this. (I generally assumed that those giving this answer did not understand this point, unless there was some indication that they did.)

4. First, the structure of these questions was intended to be "We want to know, for each part, whether there is a procedure that, for every choice of the positive integer n, answers the question." and not "The integer n is fixed at the beginning. Now we want to know, for each part, whether there is a procedure that, for that n, answers the question." In short, the procedure is to be given before the n. Each part represents an infinite number of questions (one for each n), not a single question. I think that this was fairly clear from both wording and context, although the wording could perhaps have been improved.

a. The number π is *not exactly* 22/7. Indeed, π to six digits after the decimal is 3.141592, while 22/7 is 3.142857. The latter is only a convenient approximation to the former. "Irrational" means only "cannot be written as the quotient of two integers", not "lacking rationality" or "having random digits". Thus, for example, the number .12112211122211112222... is irrational. (Challenge for doubters: Find two integers such that one divided by the other gives this number.) Neither of these points is germane to the subject matter of this course, and neither had much impact on scoring, but I thought I would mention them. Very few had any real trouble with this part.

b. There was a tendency, here again, to lapse into an answer with no explanation (e.g., merely "We do not know", which was assigned 5 points out of 10). Our doubts about their being a procedure arise not from the fact that there are an infinite number of questions (one for each n). Consider, for instance "Given positive integer n, is it prime?" This is an infinite number of questions (one for each n), and yet there certainly is a procedure for it. The doubts arise, rather, because even after we are given the n there is no sure-fire way of telling whether a "4" occurs after the n^{th} digit of π. What of the answer "There is no procedure."? The fact that we may not happen to have a procedure in hand does not mean that

none exists, only that we do not yet know. (This problem is a particularly good example of this point, for it turns out that there actually *does exist* a procedure for this particular question. But it is a subtle business.) We have as yet found no example in which we can guarantee that no procedure exists. One might even, by this point, begin to suspect that a procedure will always exist – there only remains to find it. *The entire thrust of this course is to produce an example of a question* (actually, an infinite number) *for which we can show that no procedure exists.* (Those who said "no procedure" for both parts b and c, for the same reason, were not penalized twice.)

c. It was intended that the "encoding" for this part be "our particular encoding", but it does not seem to make any real difference. The whole idea of "either... or..." was apparently confusing, and I do not think that the note on the board helped much. The question "Is n either... or..." is to have the same meaning as "Is the dog either brown or wet?" The answer is "yes" if the dog is either brown or wet, or both, and "no" otherwise. So, each n either has the property in question (either.... or...), and so for each n the answer is either "yes" or "no". But there are an infinite number of questions, namely one for each n. In particular, you are not being asked, for each n, to say which is the case (just as for the dog you are not being asked whether brownness or wetness is the case). You are not being asked for separate procedures for encoding and for primes. The issue, rather, is a procedure for the entire question. There is no simple general slogan for deciding whether there is a procedure for "or" or "and"-questions: One just has to think it through each time. (By the way, the integer "1" is not, by definition, a prime.)

The class as a whole did very well on this test. Your answers were generally clear, brief, and with good explanations. The next test will be rather like this one, but I will try to make it a little harder (not too much!), and a little more interesting. If you have any suggestion about testing in this course, I would very much appreciate them. If you feel your paper was misgraded, please follow the procedure given earlier.

Phy Sci 113 Nov 22, 1988

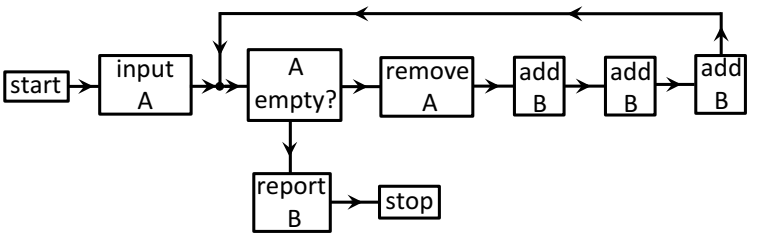
Test 2

Name:

Note: "Draw flow chart" means using only our seven instructions, with no "dashed boxes".

1. (12 points) Draw a flow chart that accepts as input nonnegative integer n; and stops if n is odd and continues forever if n is even.

2. (14 points) What does the flow chart below do? What would it do if the instruction "remove A" were replaced by "add A"?



3. (10 points) Draw the flow chart represented by the following sequence of integers: 1, 3, 1, 2, 4, 1, 3, 2.

4. (33 points) Answer ("yes", "no", or "we do not know"), and *explain* your answer *briefly*.

a. Does there exist a flow chart with no inputs that reports "0" if there exist no positive integers x, y, z, and n ≥ 3 with $x^n + y^n = z^n$; and "1" if there do exist such integers?

b. Does there exist a flow chart that accepts as input a single integer; and reports whether or not that integer is the encoding (by our encoding scheme) of some flow chart?

c. Does there exist a flow chart that accepts as input an integer that encodes some flow chart having no inputs; and reports whether or not that flow chart, when executed, will stop?

5. (24 points) Definition: A number is said to be *computable* if there exists a flow chart that accepts as input positive integer n; and reports the n^{th} digit in the decimal expansion of that number.

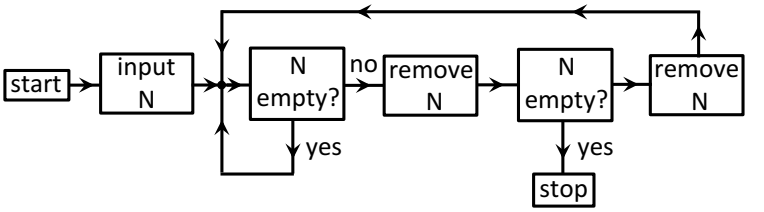
a. Draw the flow chart that shows that 1/4 is computable.

b. Give any example of a number that is not computable.

Phy Sci 113 Nov 22, 1988

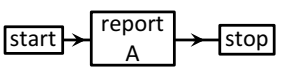
Test 2 — Solutions

1.

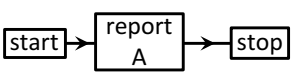


2. The flow chart accepts as input a nonnegative integer; and reports three times that integer. As modified, the flow chart would accept as input a nonnegative integer; and report "0" and stop if that integer is zero, and continue forever if that integer is other than zero.

3.



4. a. There does exist such a flow chart. Indeed, this is a single "yes-or-no" question, and so the desired flow chart is one of these:

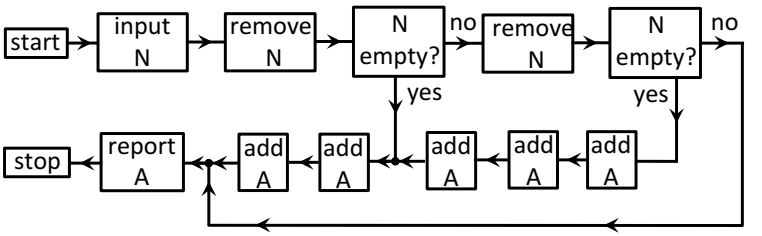


b. There does exist such a flow chart. There certainly exists a procedure to decide whether or not an integer is the encoding of some flow chart. But we have argued that every procedure can be rendered by a flow chart.

c. There does not exist such a flow chart. Indeed, assuming there existed such a flow chart, we could construct a flow chart Nancy, as follows. Let Nancy i) take the flow chart encoded at X, ii) replace its instruction "input A" by Y copies of "add A", iii) encode the resulting flow chart (which now has no input instructions) as an integer, and finally iv) run the flow chart whose existence we assumed above. Note that this would indeed be a Nancy. But our theorem says there exists no Nancy. Hence, there cannot exist a flow chart of the type described.

5. a. The flow chart must report "2" when "1" is input, "5" when "2" is input, and "0" otherwise.

b. Consider the number whose n^{th} digit is "1" if n is the encoding of a flow chart (of our special form) and A-value such that that flow chart, with that A-value, stops; and "2" otherwise. Were this number computable, then, clearly, one could construct a flow chart Nancy. But there exists no Nancy, and so this number cannot be computable.



Phy Sci 113 Nov 29, 1988

Test 2 — Comments

This test was harder than the first one. Indeed, the average score, 52, was lower than that of Test 1, although the maximum number of points, 93 was higher. Here is the distribution of scores:

Score	Nmbr receiving	Score	Numbr receiving
10-19	1	50-59	55
20-29	5	60-69	27
30-39	16	70-79	10
40-49	39	80-89	1

Comments on the individual problems:

1. This problem was basically routine, the little twist being that one had to have one's flow chart, not merely *report* the answer, but rather *do* something in response to that answer. People did well on it on the whole, with mostly minor, technical errors. For example, one is not supposed to have two flow lines entering an instruction box. (It is clear what is meant. Logically, but we have adopted the convention of one line entering a box.) One must use, instead, a node.

2. What a flow chart *does* refers to whether it stops or not, and its communications with us (i.e., its inputs and reports). Thus, for the first part of the problem, one should respond "Reports three times the integer input.", rather than "Computes three times A and stores it in B." (The latter was assigned five points out of the seven points allocated to this part of the problem.) Similarly, for the second part, what counts is that, for A greater than zero, the execution never stops, rather than that it causes bowls A and B to contain more and more beans. With these flow charts, one has to be a little careful to check the special cases, e.g., A = 0 here, and N = 0 in problem 1. It is not uncommon that a flow chart does the correct thing "in general", but fails to do so for such a special case. The class did well on this problem.

3. Virtually everyone got this easy problem. A few had not, apparently, written down on their crib sheets the rules for encoding flow charts, and there were a couple of people who felt compelled to incorporate an "input A" somewhere in the flow chart.

4. a. This is a *single* question, with answer either "yes" or "no". As such, there *must* exist a flow chart to answer it. This is the same old story we saw before, for procedures. By far the most common answer was "One can keep searching for such x, y, z, n; but cannot be sure of finding any. So, we do not know whether or not there is a flow chart." (This was assigned 3 points out of the 11 for this part.) Thus, in this sense, too, flow charts are identical to procedures. A few people had trouble with the idea of a flow chart with no "input" instruction. This is fine, as

long as the flow chart does what it is suppose to do. It is just that. Lacking an input, a flow chart can only "do" one thing, as opposed to a variety of things.

b. Here, 4 points were assigned to the answer ("yes"), the remaining 11 to the reason for that answer. It was surprisingly common for people to explain why there is a procedure for deciding whether an integer is the encoding of *other integers*, without any direct reference to whether it is the encoding of a flow chart. (This response was assigned 4 (for answer) + 2 (for reason) = 6.) Note that not every sequence of integers is the encoding of some flow chart, and, indeed, the conditions for such a sequence to be such an encoding are quite stringent. (One could, in principle, write them all out, but the result would be very long.) One should not think of there as being "two different kinds" of integers – real, ordinary integers and integers that are encodings of flow charts. An integer is an integer: It is just that one can interpret integers in various ways.

c. This was a hard problem (as promised): I think only one person saw it all the way through. The ides is that Nancy deals only with flow charts with single inputs, whereas here we are asking about flow charts with no inputs. But if one did have the flow chart asked for here, then one could use it to build a Nancy. The reason is that one can always suppress the "input" in the flow charts that Nancy examines, replacing it with a suitable number of "add" instructions. Anyway, this is a tricky point – and a good example of the kinds of arguments one makes using the "There exists no Nancy."-theorem. One shows that procedures for other things do not exist, for it if they did then one could construct a Nancy. I assigned 4 points for the answer ("no"), no matter what else was said. A few thought that the desired flow chart would have to be *exactly* a Nancy (not noticing the difference in inputs). There were quite a few papers with an intuitive sense of the answer, arguing that the desired flow chart would act much "like Nancy" (despite the difference in inputs), and so was unlikely to exist (these assigned 7 out of maximum 11). If you wish to understand how our theorem really works, understand the answer to this problem.

5. This problem, unfortunately, floundered because the definition turned out to be confusing. Apparently, there is a special style to mathematical definitions: they are hard to get if one is not used to it. It seemed, with all the explanations during the test, most people ultimately understood, basically, what the definition said. (In fact, the notion of a computable number is an important one in mathematics.)

a. Many of the errors, unfortunately, were traceable to problems with the basic definition. Thus, some wanted to "input" 1/4 at A, while others wanted to input the digits of 1/4; 2, 5, 0,... But one is supposed to input n – which digit one wants – and have the flow chart report the digit. Thus, if 1 is input, 2 is to be reported; if 2 input, 5 reported; if 3 input, 0 reported; etc. The "later" digits of .25, by the way, are "0" 's: $1/4 = .25000000000...$ A few who asked during the test were told that one counts the digits from the left (for there is no way to do it from the right) – but I do not think this was a real problem for anyone.

b. This was the second hard problem. It gave rise to an additional (unintended) issue: What, exactly, is a "number"? For our purposes, one can think of a number as being nothing more and nothing less than its digits. Thus, to "specify" a number, one need only give the *rule* for determining its digits, in succession. (Note here "rule" and not "procedure".) Thus, for example, "1/0" and "infinity" are not numbers, for they give rise to no rules. (A few people wanted to go to imaginary numbers. They won't work because, if they are to be dealt with by flow charts, one will have to encode them to integers. But, once so encoded, they

lose their apparent "noncomputability".) What the question asks, then, is for a number (i.e., a rule for the digits), without a procedure for the digits. To find one, then, one must find a "question" (which of course, will determine the digits) that has an answer, but no procedure for that answer. But we have just such a thing: the question of whether flow charts stop. So, we obtain a noncomputable number by making its digits depend on what happens to flow charts. In any case, it is very tricky, and I did not expect many to get it. Note, for example, that pi is very computable (and, indeed, we have even discussed briefly, the procedure for computing its digits). In any case, I did not expect people to find this part easy – and you didn't. A few, however, did have a general idea of how to proceed.

In terms of difficulty, etc. the final will be rather like this test.

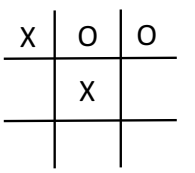
Phy Sci 113 Dec 6, 1988

Final Examination

Name:

Please answer and *briefly explain* your answer. "Does there exist flow chart?" may require answer "We don't know".

1. (20 points) consider the tic-tac-toe configuration at the right.



a. Suppose first that it is X's move. Of the five possible moves for X, which are good moves?

b. Now suppose instead that it is O's move. Is this configuration X-winnable, O-winnable, or drawable?

2. (14 points) Draw a flow chart (using only the 7 basic instructions!) that accepts integer n, and reports whether or not that integer is greater than 4.

3. (26 points) Consider the following tag game: The alphabet consists of "a", "b", and "c"; for "a" leftmost, "bc" is appended on the right, for "b", "acb", and for "c", "bac"; and in every case the leftmost three letters are deleted. We wish to construct a flow chart that will accept as input any initial string (suitably encoded), and report whether the resulting tag game will terminate, cycle, or neither.

a. describe any suitable encoding for the initial strings.

b. Does there exist such a flow chart?

4. (20 points) We wish to construct a flow chart that will decide whether or not any given candidate for a Nancy in fact works as a Nancy.

a. Now many inputs would such a flow chart have? (explain briefly.)

b. Does there exist such a flow chart?

5. (30 points) For each of the questions below, we wish to construct a flow chart to answer the question(s). How many inputs would such a flow chart have? Does there exist such a flow chart?

a. Given integer n from 1 to 9, do there exist 100 consecutive n's in the decimal expansion of π?

b. Is it the case that every integer is the difference of two primes?

c. Given integer n, do there exist positive integers x, y, and z with $x^n+y^n = z^n$?

6. (14 points) In our proof of the nonexistence of a flow chart Nancy, we showed that, given any candidate for a Nancy, there exists one particular set of

X- and Y-values for which this candidate gives the wrong answer. Show that there exists at least one other such set of values.

7. (26 points) Does there exist a flow chart that accepts, at "input X", any number that encodes a flow chart of our special form, and, at "input Y", any nonnegative integer, and

a. Stops if that flow chart with that A-value stops, and continues forever if it continues forever?

b. Stops if that flow chart with that A-value continues forever, and continues forever if it stops?

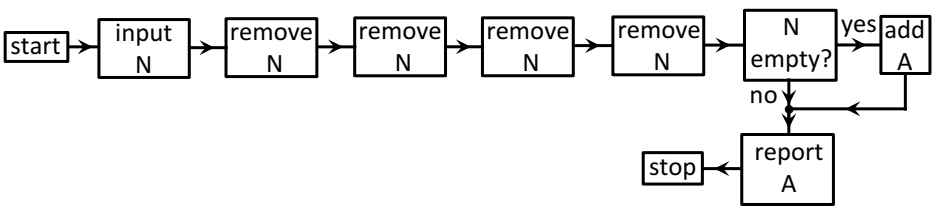
Phy Sci 113 Dec 6, 1988

Final Exam — Solutions

1. a. The move at the middle of the bottom row is not a good move. All others are. This is an X-winnable configuration, and so a move, to be a good move, need only result in an X-winnable configuration. This single move does not; and the other four do.

b. This configuration is drawable, That is, it is possible for O, against any play by X, to achieve at least a draw.

2.



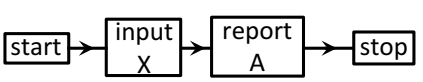
Here, "0" is reported if N is greater than four, and "1" otherwise.

3. a. We could, for example, encode an initial string by the integer having one digit for each letter in that string, where "1" stands for "a", "2" for "b", and "3" for "c". The empty string would be represented as "0". Thus, for example, initial string "bbac" would be encoded as integer "2213".

b. There does exist such a flow chart. First note that, since in no case are more than three letters appended on the right while three letters are always deleted on the left, the initial string can never grow in size. Hence, every string will either terminate or cycle: The game cannot go on forever without doing one of these two. But these two possibilities could be recognized by a flow chart. Hence, we need merely construct a flow chart that plays out the tag game, checking after each step to see whether the tag game has terminated or cycled. Eventually, one will occur.

4. a. The flow chart would have a single input. Into this input would be entered the encoding of our given candidate for Nancy.

b. There does exist such a flow chart. Indeed, here it is:



Here, our flow chart reports "0" if the candidate (encoded at X) is not a Nancy, and "1" if it is. (Recall: There exists no Nancy!)

5.a. The flow chart would have a single input, into which we are to enter the integer n (from 1 to 9). There does exist such a flow chart. There are only nine questions total, each of which has answer "yes" or "no". Hence, there is a grand total of $2^9 = 512$ possible sets of answers. For each of these, we can draw a flow chart. One (although we do not know which one!) is the correct flow chart.

b. This is a single "yes-or-no" question. There would be no inputs. There does exist such a flow chart. Indeed, here it is:

Here, "0" is reported for "no", and "1" for "yes". (Recall: The integer "19" is not the difference of two primes.)

c. There would be a single input, into which would be entered the value of n. We do not know whether or not there exists such a flow chart.

6. Fix a candidate for flow chart Nancy. The X- and Y-values we used in our proof were x= G and Y = G, where G is the integer that encodes a certain flow chart George constructed from our candidate for Nancy. Let us now construct a new flow chart, call it Henry, that is exactly the same as George, except that an extra "report S" is added just before the instruction "stop". (Almost any non-substantive change in George would do here.) Now try our candidate for Nancy again, but which X = H and Y = H, where H is the integer that encodes this flow chart Henry. But Nancy must also give the wrong answer for these X- and Y-values, by exactly the same argument as for George. Here, then, is a second such set of values. (Note that H must be a different integer from G, since Henry is a different flow chart from George.)

7. a. There does exist such a flow chart. As we have discussed, there exists a flow chart that merely "runs the flow chart encoded at X, with A-value at Y, doing whatever it would do". Such a flow chart would do nicely here.

b. There does not exist such a flow chart. Suppose, for a moment, that we had one. Then we could construct a flow chart Nancy. We would build a flow chart that first runs the flow chart answering (a) for ten steps, then the flow chart answering (b) for ten steps, then the one answering (a) for ten steps then the one answering (b) ... Eventually, either the flow chart answering (a) or the flow chart answering (b) would have to stop. By determining which one stops, and reporting the result (as "stops" if the flow chart answering (a) stops, and "continues forever" if the flow chart answering (b) stops), we would have our Nancy. But there does not exist a flow chart Nancy. Hence, there cannot exist the flow chart required for part (b).